Einführung in die Clusteranalyse mit SPSS-X für Historiker und Sozialwissenschaftler

Bacher, Johann

Historical Social Research, Vol. 14 — 1989 — No. 2, 6-167

Einführung in die Clusteranalyse mit SPSS-X für Historiker und Sozialwissenschaftler

*Johann Bacher**

Abstract: This article is addressed to users of classification procedures in the social historical sciences. According to this aim an example from historical family research is used to describe the steps necessary to solve a classification task. These steps are:

(1) Selection of classification attributes and units.
(2) Treatment of missing data.
(3) Transformation of classification attributes to comparable scales.
(4) Standardization of classification units.
(5) Selection of dissimilarity and similarity measures.
(6) Selection of classification procedures.
(7) Calculation of cluster solutions.
(8) Validition of cluster solutions by stability and sensitivity analysis.

As can be seen from the previous list some steps - especially step (2), (3) and (8) - are neglected or underestimated in most books on cluster analysis, although they are of practical importance: How can missing data be treated? What are the effects of different treatments of missing data on classification results? Is it better to transform classification attributes to comparable scales by empirical or theoretical procedures? How do these different methods of data transformation influence the results of cluster analysis? Finally, how can the validity of a cluster analysis be tested?
The article tries to answer these questions. Furthermore standard text books on cluster analysis pay little attention, how a user of statistical program packages can realize methodological rules within the program used: How can certain types of dissimilarity measures be calculated without

* Address all communications to Dr. Johann Bacher, Institute of Sociology, University of Linz, A-4020 Linz, Austria.

Historical Social Research, Vol. 14 — 1989 — No. 2, 6-167

specific option in the program used? How can data transformation be realized? Or, how can a sensitivity analysis be performed, when there is no specific program to do this?
In the article the statistical program package SPSS-X is used to demonstrate the realization of methodological rules. This investigation shows, that a wide variety of methodological rules can be realized within SPSS-X, if the user writes small programs. However there are certain limitations, especially to the treatment of missing data.
Exercises complete the represantation of the single steps. They can be solved without any computer.

1. Einleitung

1.1 Aufgabenstellung der Klassifikation und Notation

Klassifizieren und Systematisieren ist ohne Zweifel ein elmentarer Bestandteil jeder Wissenschaft. Auch für die Geschichtsforschung lassen sich ohne langes Nachdenken Beispiele für Klassifikationsprobleme finden, wie z.B. die Einteilung geschichtlicher Abläufe in Epochen, die Zuordnung archäologischer Funde zu bestimmten Epochen oder die Zuordnung von Berufen zu sozialen Schichten.

Allgemein besteht ein Klassifikationsproblem darin,

1. aus einer Menge von Objekten hinsichtlich bestimmter Merkmale homogene Klassen zu bilden oder
2. eine Menge von Objekten hinsichtlich bestimmter Merkmale bereits bekannten Klassen zuzuordnen.

Homogenität soll dann vorliegen, wenn sich die Objekte, die eine Klasse bilden, hinsichtlich der ausgewählten Merkmale ähnlich sind und sich die Klassen selbst voneinander deutlich unterscheiden. Die beiden oben dargestellten Aufgaben unterscheiden sich dadurch, daß bei der zweiten Aufgabe bereits Klassen vorliegen, die durch bestimmte Ausprägungen in den Merkmalen, die in die Klassifikation einbezogen werden, gekennzeichnet sind. Bei der ersten Aufgabe sollen diese Klassen erst bestimmt werden. In dieser Arbeit werden vor allem Aufgaben der ersten Art behandelt. Verfahren, die die zweite Aufgabe zu lösen versuchen, wie z.B. die Diskriminanzanalyse, werden nur insofern besprochen, als sie sich zur Überprüfung der Ergebnisse der ersten Aufgabe eignen.

Formal werden für die Lösung einer Klassifikationsaufgabe eine Menge von Objekten, eine Menge von Merkmalen, in denen die Objekte Ausprägungen besitzen und die die gesuchten Klassen charakterisieren, sowie ein geeignetes Verfahren benötigt. Für die weiteren Ausführungen wird, da die Terminologie in der Klassifikationsliteratur keinesfalls einheitlich ist, folgende Notation festgesetzt:

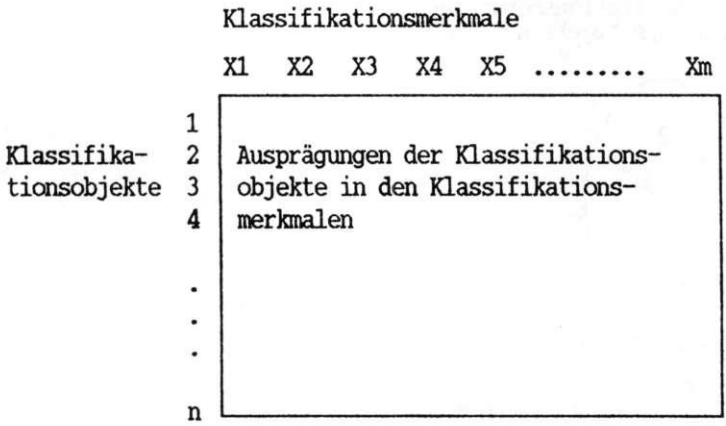
- Die zu klassifizierenden Objekte werden als **Klassifikationsobjekte** bezeichnet. In der Literatur verwendete synonyme Bezeichnungen sind (Vogel 1975: 45): Elemente (»elements«), Fälle (»cases«), OTUs (»operational taxonomic units«), Items u.a..
- Die Merkmale, die in die Klassifikation eingehen, werden als **Klassifikationsmerkmale** bezeichnet. Synonyme Bezeichnungen sind (Vogel 1975: 48): Variable (»variables«, »variates«), Charakteristiken (»characteristics«, »characters«), Attribute (»attributes«), Muster (»patterns«, »features«) u.a..
- Eine **empirische Klassifikation** soll dann vorliegen, wenn den Klassifikationsobjekten und - merkmalen empirische Beobachtungen zugrundeliegen. Synonyme Ausdrücke sind (Vogel 1975: 1): Numerische - oder mathematische Taxonomie (»numerical« oder »mathematical taxonomie«), Mustererkennung (»pattern recognition«) u.a..
- Mathematisch - statistische Verfahren, die eine empirische Klassifikation durchführen, sollen als **empirische Klassifikationsverfahren** bzw. oft kurz als Klassifikationsverfahren bezeichnet werden. Für sie werden oft die gleichen Synonyme verwendet wie für die empirische Klassifikation selbst.
- Die Ergebnisse einer Klassifikation schließlich werden als **Klassen** oder **Cluster** bezeichnet. Sic werden oft auch Typen oder Gruppen genannt.

Die Klassifikationsobjekte und - merkmale müssen keineswegs mit den zur Verfügung stehenden empirischen Beobachtungen identisch, sie müssen aber bei einer empirischen Klassifikation zumindest aus empirischen Beobachtungen abgeleitet sein. So kann beispielsweise die Geburtenrate als Klassifikationsmerkmal verwendet werden, empirisch beobachtet werden aber nur die Anzahl der Geburten in einem Jahr und die Bevölkerungsgröße zu Beginn und am Ende des Jahres (1). Klassifikationsobjekte können z.B. auch durch Aggregation der beobachteten Einheiten entstehen.

Klassifikationsobjekte und - merkmale bilden eine **Klassifikationsdatenmatrix.** Die Abbildung 1.1-1 enthält den allgemeinen Aufbau einer solchen Matrix. Die Zeilen der Klassifikationsdatenmatrix werden durch die Klassifikationsobjekte 1,2,..., n gebildet, die Spalten durch die Klassifikationsmerkmale XI, X2, X3, ..., Xm. In der Matrix stehen die Ausprägungen der Klassifikationsobjekte in den Klassifikationsmerkmalen.

Abbildung 1.1-1:

Aufbau einer Klassifikationsdatenmatrix



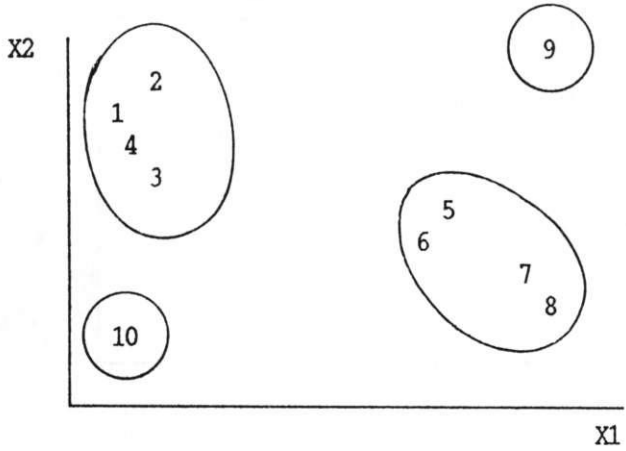
Geometrisch kann man sich die Klassifikationsdatenmatrix als m - dimensionalen Raum, der von den Klassifikationsmerkmalen aufgespannt wird, vorstellen. In diesem Raum besitzt jedes Klassifikationsobjekt eine bestimmte Lage (vgl. Abbildung 1.1-2).

In der Abbildung erkennt man unmittelbar 4 Cluster: Ein Cluster wird durch die Klassifikationsobjekte 1, 2, 3 und 4 gebildet, ein weiteres durch die Klassifikationsobjekte 5, 6, 7 und 8. Die Klassifikationsobjekte 9 und 10 bilden jedes für sich ein selbständiges Cluster. Die Klassifikationsmerkmale müssen keinesfalls - wie in der Abbildung 1.1-2 - unabhängig voneinander sein. Liegt keine Unabhängigkeit vor, stehen die Klassifikationsmerkmalc nicht mehr rechtwinkelig (orthogonal) aufeinander. Dieses Problem korrelierter (nicht unabhängiger) Klassifikationsmerkmale wird in Abschnitt 2.3 behandelt.

In dem Beispiel der Abbildung 1.1-2 kann ohne Zuhilfenahme eines mathematischen Verfahrens eine Klassifikation vorgenommen werden. Bei einer größeren Anzahl von Klassifikationsmerkmalen und/oder einer größeren Objektmenge ist das nicht mehr bzw. äußerst schwer möglich. In diesem Fall ist man auf die Anwendung empirischer Klassifikationsverfahren angewiesen. Wie wird nun eine empirische Klassifikationsaufgabe konkret gelöst?

Abbildung 1.1-2:

Geometrische Darstellung einer Klassifikationsdatenmatrix mit 2 Klassifikationsmerkmalen X1 und X2 und n=10 Klassifikationsobjekten



1.2 Lösungsschritte einer empirischen Klassifikationsaufgabe

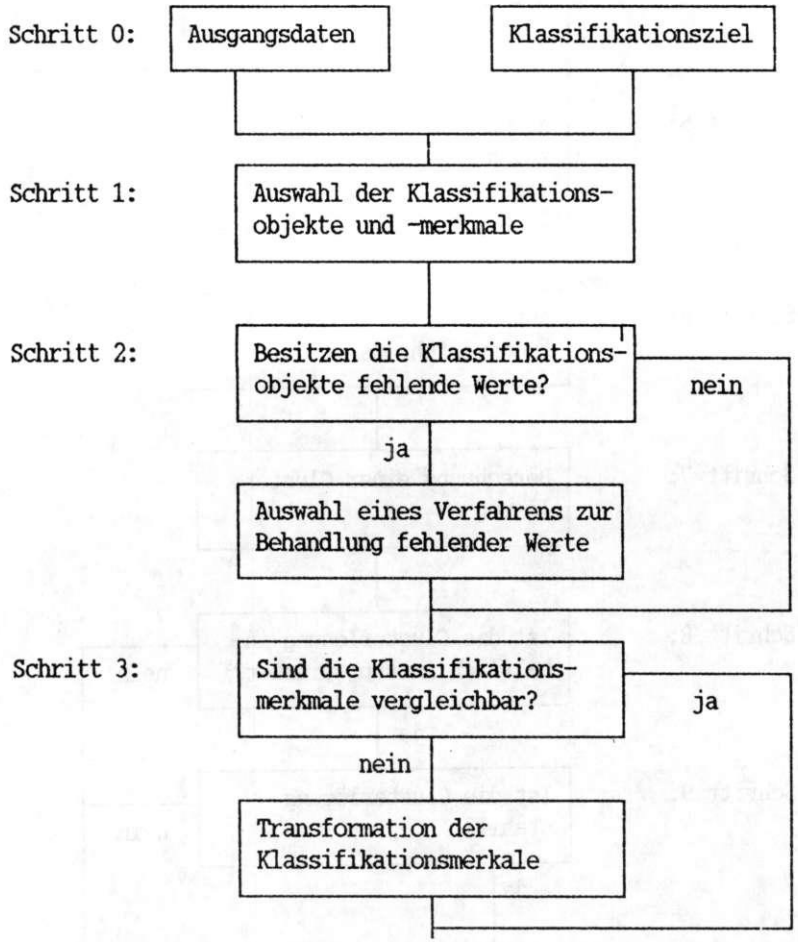
Die Abbildung 1.2-1 enthält ein allgemeines Schema für die Lösung einer Klassifikationsaufgabe. Gegeben sind ein Klassifikationsziel, z.B. die Bildung sozialer Klassen, und Ausgangsdaten, z.B. die Ergebnisse einer Volkszählung. Den ersten Schritt stellt die **Auswahl der Klassifikations-merkmale und - objekte** aus den zur Verfügung stehenden Ausgangsdaten dar (s. dazu Abschnitt 2.2). Ideal wäre selbstverständlich, wenn die Da-tensammlung selbst von dem Klassifikationsziel gesteuert würde, was aber sehr häufig nicht der Fall ist.

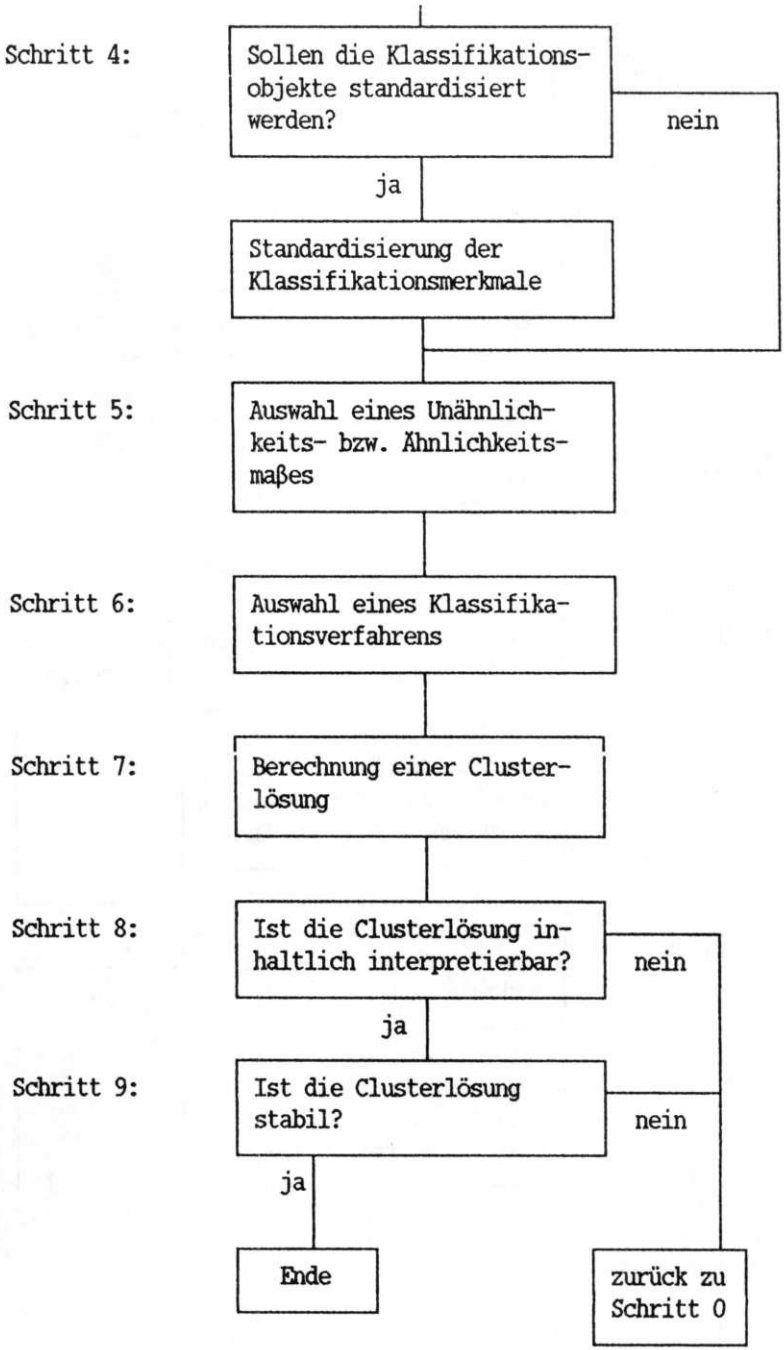
Im nächsten Schritt ist zu überprüfen, ob **fehlende Werte** vorliegen und wie sie gegebenenfalls zu behandeln sind (s. dazu Abschnitt 4.1).

Die **Vergleichbarkeit** der Klassifikationsmerkmale wird im nächsten Schritt bei der Lösung einer Klassifikationsaufgabe überprüft (s. dazu Ab-schnitt 2.3). Ist eine Vergleichbarkeit nicht gegeben, muß eine **Transfor-mation der Klassifikationsmerkmale** durchgeführt werden (s. dazu Ab-schnitt 4.2).

Abbildung **1.2-1:**

Lösungsschema einer Klassifikationsaufgabe




Schritt 4:
Sollen die Klassifikations-objekte standardisiert werden?
nein
ja
Standardisierung der Klassifikationsmerkmale
Schritt 5:
Auswahl eines Unähnlich-keits- bzw. Ähnlichkeits-maßes
Schritt 6:
Auswahl eines Klassifika-tionsverfahrens
Schritt 7:
Berechnung einer Cluster-lösung
Schritt 8:
Ist die Clusterlösung in-haltlich interpretierbar?
nein
ja
Schritt 9:
Ist die Clusterlösung stabil?
nein
ja
Ende
zurück zu Schritt 0

In den nächsten Schritten muß entschieden werden, ob die **Klassifikationsobjekte standardisiert** werden und welches **Maß zur Messung der Unähnlichkeit bzw. Ähnlichkeit** der Klassifikationsobjekte verwendet wird (s. dazu Abschnitt 2.4.2). Eine Standardisierung der Klassifikationsobjekte wird beispielsweise dann vorgenommen, wenn nur relative Unterschiede zwischen den Klassifikationsobjekten in die Analyse einbezogen werden sollen (s. dazu Abschnitt 2.4.3). Die Entscheidung für ein bestimmtes Unähnlichkeits - bzw. Ähnlichkeitsmaß hängt u.a. formal von dem **Meßniveau der Klassifikationsmerkmale** ab (s. dazu Abschnitt 1.3).

Der nächste Schritt besteht in der **Auswahl eines Klassifikationsverfahrens.** Diese lassen sich grob in drei Gruppen einteilen:

- Verfahren der Clusteranalyse,
- mehrdimensionale Skalierungsverfahren sowie
- Hauptkomponenten- und Faktorenanalyse.

Die mehrdimensionale Skalierung sowie die Hauptkomponenten- und Faktorenanalyse stellen unvollständige Klassifikationsverfahren dar (Mezzieh 1982: 154), da sie nur den Klassifikationsmerkmalsraum reduzieren und eine graphische Darstellung der Klassifikationsmerkmale und Klassifikationsobjekte in diesem reduzierten Raum ermöglichen. Die Klassifikation selbst wird i.d.R. dann durch eine graphische Inspektion vorgenommen. In der Forschungspraxis besteht der Unterschied zwischen der Clusteranaylse und den anderen Verfahren (mehrdimensionale Skalierung, Faktoren- und Hautpkomponentenanalyse) darin, daß die Clusteranalyse zur Analyse der Klassifikationsobjekte eingesetzt wird, die anderen Verfahren dagegen für eine Analyse der Klassifikationsmerkmale. Die Verfahren selbst zwingen dieses Vorgehen nicht auf. Die Verfahren der Clusteranalyse können prinzipiell auch für eine Analyse der Klassifikationsmerkmale (Abschnitt 2.6) eingesetzt werden.

Nach der Berechnung einer Clusteranalyse muß diese auf ihre Gültigkeit hin überprüft werden. Voraussetzung für eine gültige Clusterlösung ist ihre **inhaltliche Interpretierbarkeit** und **ihre Stabilität.** Verfahren der Stabilitätsprüfung werden ausführlich in Kapitel 3 behandelt. Ist die Gültigkeit einer Clusterlösung nicht gegeben, wird zum Schritt 0 zurückgekehrt. Die getroffenen Entscheidungen werden schrittweise überprüft und gegebenenfalls inhaltlich begründet modifiziert. Ein mehrmaliges Durchgehen der Schritte ist der Regelfall.

13 Exkurs: Das Meßniveau von Klassifikationsmerkmalen

Es lassen sich drei Arten von Meßniveaus unterscheiden:

- **nominales Meßniveau:** Zwischen den Merkmalsausprägungen besteht nur eine gleich/ungleich-Relation. Die Stellung einer Person in einem Haushalt mit den Ausprägungen

  Haushaltsvorstand
  Ehegattin des Haushaltsvorstandes
  Sohn des Haushaltsvorstandehepaares
  Tochter des Haushaltsvorstandehepaares
  Verwandte des Haushaltsvorstandehepaares
  Dienstboten
  Bettgeher

  ist ein Beispiel für ein nominales Klassifikationsmerkmal. Nominales Meßniveau ermöglicht also nur eine Einteilung in Klassen (Ausprägungen) und die Feststellung von Ungleichheit (zwei Klassifikationsobjekte besitzen unterschiedliche Ausprägungen) bzw. von Gleichheit (zwei Klassifikationsobjekte besitzen die gleiche Ausprägung).

- **ordinales Meßniveau:** Zwischen den Merkmalsausprägungen besteht eine Ordnungsrelation (gleich/kleiner/größer-Relation). Die Größe der landwirtschaftlichen Nutzfläche mit den Ausprägungen

  1 = 0 bis 1 ha
  2 = 1 bis 2 ha
  3 = 2 bis 5 ha
  4 = 5 und mehr ha

  ist ein Beispiel für ein ordinales Klassifikationsmerkmal, da sich beispielsweise angeben läßt, daß ein landwirtschaftlicher Betrieb mit der Ausprägung 4 (5 und mehr ha) über eine größere landwirtschaftliche Nutzfläche verfügt als ein Betrieb mit der Ausprägung 2 (1 bis 2 ha). Neben der Gleichheit und Ungleichheit kann bei ordinalem Meßniveau noch eine Größenordnung festgestellt werden.

- **quantitatives Meßniveau:** Zwischen den Merkmalsausprägungen sind Abstände definiert. Würde die landwirtschaftliche Nutzfläche - ohne eine Klasseneinteilung - direkt in Hektar erfaßt, würde ein quantitatives Klassifikationsmerkmal vorliegen. Es läßt sich nun nicht nur feststellen, daß ein Betrieb über mehr landwirtschaftliche Nutzfläche verfügt als ein anderer, sondern auch um wieviel mehr. Das quantitative Meßniveau kann noch in ein intervall - und ratioskaliertes

Meßniveau unterteilt werden. In der Praxis kommt dieser Differenzierung aber keine Bedeutung zu.

Für nominales Meßniveau ist auch die Bezeichnung kategoriales Meßniveau üblich. Nominales und ordinales Meßniveau werden ferner oft noch als nicht-metrisches Meßniveau bezeichnet. Für »quantitativ« wird oft auch die Bezeichnung »metrisch« oder »kontinuierlich« verwendet.

Die Information, die Klassifikationsmerkmale enthalten, nimmt vom nominalen Meßniveau ausgehend über das ordinale Meßniveau bis zum quantitativen Meßniveau kontinuierlich zu. Deshalb wird auch von einer Hierarchie des Meßniveaus gesprochen, mit dem nominalen Meßniveau am unteren Ende der Hierarchie und dem quantitativen bzw. ratioskalierten am oberen Ende.

Übungsaufgabe 1: Bestimmen Sie das Meßniveau der folgenden Klassifikationsmerkmale und begründen Sie Ihre Entscheidung!

a) Klassifikationsmerkmal - »Anzahl der im Haushalt lebenden Kinder« mit den Ausprägungen 0, 1,2, 3, ..
b) Klassifikationsmerkmal - »Familienstand« mit den Ausprägungen »ledig«, »verheiratet«, »verwitwet« und »geschieden«.
c) Klassifikationsmerkmal — »Betriebsgröße« mit den Ausprägungen »1« (1 - 10 Beschäftigte), »2« (11 - 50 Beschäftigte), »3« (51 - 100 Beschäftigte) und »4« (mehr als 100 Beschäftigte).
d) Klassifikationsmerkmal = »Industrialisierungsgrad« gemessen als Anteil des industriellen Sektors am Bruttoinlandsprodukt.

1.4 Ein Wegweiser durch das Skript

In diesem Skript werden unter den Verfahren der Clusteranalyse nur die sogenannten hierarchisch - agglomerativen Verfahren (s. dazu Abschnitt 2.4.1) und ein allokatives Verfahren (s. dazu Abschnitt 4.1.2 und 7.2) behandelt. Über weitere Verfahren informiere man sich beispielsweise bei Kaufmann und Pape (1984). Die Begründung für diese Auswahl ist eine pragmatische: Die ausgewählten Verfahren stehen in dem Statistikprogrammpaket SPSS X zur Verfügung, das eine sehr große Verbreitung besitzt und für das an den meisten Universitäten einführende Lehrveranstaltungen angeboten werden. Das vorliegende Skript setzt diese Grundkenntnisse voraus. Das Schwergewicht liegt auf der konkreten Umsetzung und technischen Realisierung methodologischer Regeln, da dies in den meisten einführenden Lehrbüchern zur Clusteranalyse nicht behandelt wird. Gerade hier liegen aber sehr oft die eigentliche Probleme. Diese lassen sich aber nur demonstrieren, wenn mit einem bestimmten Programmsystem gearbeitet wird.

In Kapitel 2 wird deshalb ein Anwendungsbeispiel einer hierarchischen Clusteranalyse ausführlich dargestellt. Die Schritte 2 und 3 (Behandlung fehlender Werte und Transformation Klassifikationsmerkmalen) des Schemas der Abbildung 1.2-1 werden dabei übersprungen. Leser, die mit der hierarchischen Clusteranalyse vertraut sind, können unmittelbar mit Kapitel 3 und 4 beginnen.

In diesen beiden Kapitel werden Verfahren der Stabilitätsprüfung (Kapitel 3) und Verfahren der Behandlung fehlender Werte (Kapitel 4) behandelt.

Jedes Kapitel ist in Abschnitte unterteilt, wobei jeder Abschnitt mit Übungsaufgaben abgeschlossen wird. Für deren Lösung ist nur ein Blatt Papier und eventuell ein Taschenrechner erforderlich. »Trockenschwimmkurse« dieser Art werden heute vielfach kritisiert und als überflüssig empfunden. Unmittelbares Ausprobieren an einem Computer scheint inzwischen zu einem didaktischen Muß geworden zu sein. Der Autor vertritt dagegen die Auffassung, daß zum Verständnis der Logik statistischer Verfahren das (händische) Durchrechnen von einfachen Beispielen entscheidend beitragen kann. Darüber hinaus ist es auch bei der Benutzung von SPSS-X oder anderen Statistiksoftwaresystemen vorteilhaft, die Programme zunächst auf einem Blatt Papier niederzuschreiben, insbesondere wenn komplexere Operationen erforderlich sind. Wer an einem überlastetet Rechenzentrum arbeitet oder gearbeitet hat, wird sich dieser Meinung ohne Zweifel anschließen.

2. Hierarchische Clusteranalyse: Ein Anwendungsbeispiel

2.1 Klassifikationsziel und Ausgangsdaten

Die in Kapitel 1 dargestellten Schritte, die bei einem Klassifikationsproblem zu lösen sind, sollen nun anhand eines Beispiels erörtert werden. Das Ziel der Klassifikation besteht darin, familiale Haushaltsstrukturen in einer osttiroler Gemeinde am Ende des 18. Jahrhunderts zu bestimmen. Die Daten wurden von Ursula Walter erhoben (2). Sie sollen im folgenden nur so weit beschrieben werden, als es für das Verständnis der weiteren Ausführungen notwendig ist. Zur Beschreibung der Ausgangsdaten werden für dieses und für folgende Beispiele nachstehende Begriffe verwendet:

Die Einheiten, für die Informationen erhoben wurden, werden als **Objekte, Einheiten** oder **Personen** bezeichnet. Für die erhobenen Informationen wird die Bezeichnung **Variablen** oder **Merkmale** verwendet. Die Gesamtheit der Information eines Objektes - die Ausprägungen in den Variablen - wird als Datensatz bezeichnet, die Gesamtheit der Information in allen Objekten schließlich als Datenmatriv Nochmals sei ausdrücklich betont, daß diese Datenmatrix keinesfalls mit der Klassifikationsdatenmatrix identisch sein muß.

Charakteristisches formales Merkmal dieser Ausgangsdaten ist, daß jede Person, die in einem Haushalt lebt, einen Datensatz bildet. Für jede Person wurden u.a. folgende Variablen erhoben:

- Die Variable HNR **»Hausnummer«** mit den Ausprägungen 1, 2, 3, ...
- Die Variable HHNR **»Haushaltsnummer«** mit den Ausprägungen 1, 2, 3, ... Diese Variable wurde eingeführt um Haushalte identifizieren zu können, die in einem Haus, z.B. einem Mietshaus, wohnen. Jeder Haushalt (HH) ist also eindeutig durch die beiden Identifikationsvariablen »Hausnummer« und »Haushaltsnummcr« gekennzeichnet.
- Die Variable RU **»Stellung im Haushalt«.** Diese Variable umfaßte ursprünglich 26 Ausprägungen (3), die für die weitere Analyse zu 4 Ausprägungen zusammengefaßt wurden. Diese 4 Ausprägungen sind: Mitglied der Kernfamilie (=1), Mitglied der Verwandtschaft der Kernfamilie (=2), Inwohner (=3) und Gesinde (=4). Zur Kernfamilie wurden dabei gezählt: Der Haushaltsvorstand, dessen Ehefrau und die Kinder des Haushaltsvorstandehepaares. Als Verwandte wurden bezeichnet: Die Eltern und Geschwister des Haushaltsvor-

Standehepaares, die Kinder dieser Geschwister, Schwiegerkinder und mögliche Enkelkinder. Zu der Kategorie »Inwohner« wurden die Inwohner selbst und deren Kinder zusammengefaßt. Die zusammengefaßte Kategorie »Gesinde« enthält Hilfskräfte, das Gesinde und deren Kinder.

Die Tabelle 2.1-1 enthält die Ausprägungen in diesen Variablen für die ersten 15 Personen. Für die Variable »Stellung im Haushalt« wurden dabei die Ausprägungen vor und nach der Zusammenfassung angegeben.

Tabelle 2.1-1:

Struktur der Ausgangsdaten

Erfaßte Einheit:	Variablen: HNR	HHNR	RU(vor d. Zs.fassung)	Ausprägung	RU(nach d. Zs.fassung)
1	1351	1	1	Haushaltsvorstand (HV)	1
2	1351	1	15	Hilfskraft	4
3	1351	1	12	Gesinde	4
4	1351	1	12	Gesinde	4
5	1381	1	1	Haushalts vorstand (HV)	1
6	1381	1	2	Ehefrau des HV	1
7	1381	1	3	Kind des HVehepaares	1
8	1381	1	11	Ehegatte des Kindes des HVehepaares	2
9	1381	1	6	Kind d. Sohnes/Tochter des HVehepaares	2
10	1381	1	6	" "	2
11	1381	1	13	Inwohner	3
12	1411	1	1	Haushalts vorstand (HV)	1
13	1411	1	2	Ehefrau des HV	1
13	1411	1	3	Kind des HVehepaares	1
13	1411	1	3	" "	1
14	1411	1	3	" "	1
15	1421	1	1	Haushaltsvorstand (HV)	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Der Haushalt mit der Hausnummer 1351 und der Haushaltsnummer 1 besteht aus einem Haushaltsvorstand, einer Hilfskraft und zwei Mitgliedem, die dem Gesinde angehören. Eine vollständige Kernfamilie, die zumindest aus einem Ehepaar und einem Kind besteht (Zweigenerationen-

familie), liegt bei diesem Haushalt nicht vor. Tatsächlich handelt es sich bei diesem Haushalt um den Haushalt des örtlichen Pfarrers. Für die Clusteranalyse wurde dieser Fall nicht eliminiert, aber wir können bereits an dieser Stelle die Forderung an ein brauchbares Ergebnis der Clusteranalyse formulieren, nämlich, daß Haushalte mit dieser oder einer ähnlichen Struktur als selbständiges Cluster identifiziert und eindeutig von anderen Clustern mit einer vollständigen Kernfamilie getrennt werden müßten.

In dem zweiten Haushalt wohnt das Haushaltsvorstandehepaar mit einem Kind, mit dessen Ehegatten und zwei Enkelkindern, also eine Dreigenerationenfamilie, sowie ein Inwohner. Der dritte in der Tabelle vollständig angeführte Haushalt besteht schließlich aus dem Haushaltsvorstandehepaar und deren drei Kinder, also nur aus Mitgliedern der Kernfamilie.

In der Tabelle 2.1-1 ist ferner deutlich ersichtlich, daß die von uns durchgeführte Zusammenfassung nicht ein bestimmtes Ausmaß der Willkürlichkeit entbehrt, insbesondere in bezug auf die Abgrenzung von Kernfamilie und Verwandten.

2.2 Auswahl der Klassifikationsmerkmaie und -Objekte

Durch das Klassifikationsziel sind die Haushalte als Klassifikationsobjekte definiert. Für die Klassifikationsmerkmale wird die inhaltliche Forderung gestellt, daß sie die familiale Struktur in einem Haushalt abbilden sollen. Im Detail wurden für das Beispiel folgende Klassifikationsmerkmale ausgewählt:

- die Anzahl (AKERNF) der im Haushalt lebenden Mitglieder der Kernfamilie,
- die Anzahl (AVERW) der im Haushalt lebenden Verwandten,
- die Anzahl (AINW) der im Haushalt lebenden Inwohner und
- die Anzahl (AGESIN) des im Haushalt lebenden Gesindes.

Diese Klassifikationsmerkmale können aus der zusammengefaßten Variablen »Stellung im Haushalt« gewonnen werden. Die Struktur der gesuchten Klassifikationsdatenmatrix ist in der Abbildung 2.2-1 dargestellt. Sie enthält als Zeilen die Haushalte, als Spalten die Klassifikationsmerkmale AKERNF, AVERW, AINW und AGESIN und als Elemente die Ausprägungen der Haushalte in diesen Klassifikationsmerkmalen.

Um diese Matrix zu erhalten, müssen lediglich in jedem Haushalt die Anzahl der Mitglieder der Kernfamilie, der Verwandten, der Inwohner und des Gesindes gezählt werden. Diese Operation wird als **Aggregation** bezeichnet: Die Personen eines Haushaltes und ihre Merkmalsausprägungen werden zu einem Haushalt aggregiert. Technisch läßt sich dieser Vorgang in SPSSX folgendermaßen durchführen:

```
TITLE »Analyse der familialen HH-typen«
FILE HANDLE FAMDAT / NAME=»FAM.DAT«
GET FILE FAMDAT
RECODE RU (1 THRU 5=1) (6,7,8,9,10,21,24,25,26=2)
          (7,13,22=3) (12,14,15,23=4) (ELSE=0)
COMPUTE KERNF = 0
IF (RU = 1) KERNF = 1
COMPUTE VERW = 0
IF (RU = 2) VERW = 1
COMPUTE INW = 0
IF (RU = 3) INW = 1
COMPUTE GESIN = 0
IF (GESIN = 4) GESIN = 1
AGGREGATE OUTFILE = *
  /BREAK = HNR HHNR
  /AKERNF AVERW AINW AGESIN = SUM(KERNF VERW INW GESIN)
```

Abbildung 2.2-1:

Struktur der gewünschten KLassifikationsdatenmatrix

Klassifikations-objekte:	Klassifikationsmerkmale: AKERNF AVERW AINW AGESIN
HH 1	
HH 2	Ausprägungen der Haushalte in den Klassifikationsmerkmalen
HH 3	
.	
.	
.	

Durch die Anweisung TITLE **»Analyse der familialen HHtypen«** wird dem SPSS-X Programm ein Name gegeben, der auf jeder Seite des Ergebnisausdruckes steht und vorwiegend Dokumentationszwecken dient.

Durch die Anweisung FILE HANDLE FAMDAT wird die SPSS X Arbeitsdatei FAMDAT definiert. Der Befehl NAME - »FAM.DAT« legt die Beziehung zwischen dem Namen der SPSS-X internen Arbeitsdatei und dem Namen FAM.DAT, unter dem die Datei extern abgespeichert ist, fest. Die FILE HANDLE Anweisung ist von dem zur Verfügung stehenden

Betriebssystem abhängig (4). In zahlreichen Betriebssystemen ist die Angabe des externen Dateinamens nicht erforderlich, es genügt also z.B. der Befehl FILE HANDLE FAMDAT, wobei der Name der SPSS-X internen Arbeitsdatei mit dem externen Dateinamen übereinstimmen muß.

Der Befehl GET FILE FAMDAT bewirkt, daß die Datei mit dem Namen FAM.DAT als SPSS-X interne Arbeitsdatei »geladen« wird.

Sie steht nun für Datenmanipulationen und statistische Auswertungen zur Verfügung.

Durch die RECODE - Anweisung wird die Zusammenfassung der Variablen RU (Stellung im Haushalt) in die vier Ausprägungen »Mitglied der Kernfamilie« (=1), »Mitglied der Verwandten« (=2), »Inwohner« (=3) und »Gesinde« (=4) durchgeführt. Alle anderen Ausprägungen - einschließlich fehlender Werte - erhalten durch den ELSE - Befehl den Wert 0.

Die Variable »Stellung im Haushalt« besitzt nur nominales Meßniveau. »Zahlen« im Sinne von »Addieren« ist nun aber eine Operation, die nur für quantitative Variablen definiert ist. Durch die Auflösung der nominalen Variablen RU in sogenannte **DummyVariablen** (»Scheinvariablen«) kann diese Anwendungsvoraussetzung erfüllt werden. Diese Auflösung wird durch die folgenden COMPUTE - und IF - Befehle erreicht.

Die Anweisung COMPUTE KERNF - 0 bewirkt, daß der Dummy-Variablen KERNF der Wert 0 zugewiesen wird. Die daran anschließende Anweisung IF (RU = 1) KERNF = 1 führt dazu, daß jene Personen, die der Kernfamilie angehören, in der Dummy-Variablen KERNF den Wert 1 erhalten. Derselbe Vorgang wird für die anderen drei Ausprägungen der zusammengefaßten Variablen RU durchgeführt. Die Anweisung COMPUTE VERW = 0 bewirkt, daß die Dummy-Variable VERW zunächst den Wert 0 erhält. Die daran anschließende Anweisung IF (RU = 2) VERW = 1 führt dazu, daß jene Personen, die der Verwandtschaft der Kernfamilie angehören, in der Dummy Variablen VERW den Wert 1 erhalten, usw...

SPSS-X intern führen diese Anweisungen zur Erweiterung der Arbeitsdatei mit den Dummy-Variablen KERNF, VERW, INW und GES1N (s. Tabelle 2.2-1). In dieser erweiterten Datei müssen nur mehr die Dummy-Variablen für jeden Haushalt aufsummiert werden, um die gesuchten Klassifikationsmerkmale und - Objekte zu erhalten.

Diese Summation kann mit der SPSS-X Prozedur AGGREGATE (SPSS Inc. 1986: 248 - 259) durchgeführt werden, die durch den Befehl AGGREGATE aufgerufen wird.

Die Anweisung OUTFILE = * bewirkt, daß die Ergebnisse der Aggregation als neue SPSS-X Arbeitsdatei angelegt werden. Die der AGGREGATE-Prozedur folgenden Befehle beziehen sich auf diese neue Arbeitsdatei. Die ursprüngliche Arbeitsdatei steht dann nicht mehr zur Verfügung. Durch die Anweisung BREAK - HNR HHNR wird festgelegt, daß

Tabelle 2.2-1:

Struktur der neuen SPSS-X Arbeitsdatei

Erfaßte Ein - heit:	Variablen d. Ausgangsdaten: HNR	HHNR	RU(nach d. Zs.fassung)	Dummy-Variablen: KERNF	VERW	INW	GESIN
1	1351	1	1	1	0	0	0
2	1351	1	4	0	0	0	1
3	1351	1	4	0	0	0	1
4	1351	1	4	0	0	0	1
				-	-	-	-
			(Σ =	1	0	0	3)
5	1381	1	1	1	0	0	0
6	1381	1	1	1	0	0	0
7	1381	1	1	1	0	0	0
8	1381	1	2	0	1	0	0
9	1381	1	2	0	1	0	0
10	1381	1	2	0	1	0	0
11	1381	1	3	0	0	1	0
				-	-	-	-
			(Σ =	3	3	1	0)
12	1411	1	1	1	0	0	0
13	1411	1	1	1	0	0	0
13	1411	1	1	1	0	0	0
13	1411	1	1	1	0	0	0
14	1411	1	1	1	0	0	0
				-	-	-	-
			(Σ =	5	0	0	0)
15	1421	1	1	1	0	0	0
.	.	.	.	.			.
.	.	.	.	.			.
.	.	.	.	.			.

über die Variablen HNR und HHNR, die einen Haushalt identifizieren, aggregiert wird. Die Anweisung AKERNF AVERW AINW AGESIN = SUM(KERNF VERW INW GESIN) bewirkt, daß die Summe der Dummy-Variablen KERNF, VERW, INW und GESIN berechnet und den neuen Variablen AKERNF, AVERW, AINW und AGESIN zugewiesen wird. Die Zuweisung findet in derselben Reihenfolge statt, wie die Dummy-Variablen in der Summenfunktion aufgelistet sind. Die Summe der Dummy-Variablen KERNF für einen Haushalt wird also der neuen Variablen AKERNF zugewiesen, die Summe der Dummy-Variablen VERW der neuen Variablen AVERW, usw..

Die so erzeugte neue SPSS·X Arbeitsdatei ist die von uns gesuchte Klassifikationsdatenmatrix. Die Ausprägungen der ersten fünf Haushalte in den Klassifikationsmerkmalen sind in der Tabelle 2.2-2 abgebildet.

Tabelle 2.2-2:

Die Ausprägungen der ersten fünf Klassifikationsobjekte der gewünschten Klassifikationsdatenmatrix

Klassi - fikations -	Klassifikationsmerkmale:			
Objekte:	AKERNF	AVERW	AINW	AGESIN
HH 1	1.00	0.00	0.00	3.00
HH 2	3.00	3.00	1.00	0.00
HH 3	5.00	0.00	0.00	0.00
HH 4	3.00	2.00	0.00	0.00
HH 5	3.00	0.00	0.00	0.00

Hinweise zur SPSS·X Prozedur AGGREGATE:

1.) Die Ergebnisse der AGGREGATE·Prozedur können durch folgende Anweisungen extern zwischengespeichert werden:

```
FILE HANDLE AFAMD / NAME - »AFAM.DAT«
AGGREGATE OUTFILE - AFAMD
          /...
```

Durch die FILE HANDLE·Anweisung wird wiederum die Beziehung zwischen interner SPSS·X Datei und der externen Speicherdatei hergestellt. Die aggregierten Daten werden durch die Anweisung OUTFILE = AFAMD auf die externe Datei mit dem Namen AFAM.DAT gespeichert. Die Speicherung wird i.d.R. nur lokal durchgeführt und kann durch entsprechende Betriebssystemanweisungen permanent gesichert werden. Diese Befehle sind - wie die FILE HANDLE Anweisung - vom zur Verfügung stehenden Betriebssystem abhängig und können hier nicht im Detail besprochen werden.

2.) Es ist oft sinnvoll, insbesondere bei komplexen mathematischen Operationen, die Ergebnisse dieser Operationen ausdrucken zu lassen. In unserem Beispiel kann dafür folgende Anweisung verwendet werden:

```
LIST VARIABLES = HNR HHNR RU KERNF VERW INW GESIN /
    CASES FROM 1 TO 20
```

Diese Anweisung muß vor dem Befehl AGGREGATE stehen und bewirkt die Ausgabe der ersten 20 Fälle.

Die LIST-Anweisung kann auch für eine Ausgabe der Aggregationsergebnisse benutzt werden. Die entsprechende Anweisung ist:

```
LIST VARIABLES = HNR HHNR AKERNF AVERW AINW AGESIN /
    CASES FROM 1 TO 5
```

und muß nach der AGGREGATE-Prozedur stehen.

3.) Neben der Summenfunktion stehen in SPSS-X u.a. noch folgende Funktionen zur Verfügung:

MEAN(Variablenliste) - anstelle von Summen werden Mittelwerte berechnet
SD (Variablenliste) - anstelle von Summen werden Standardabweichungen berechnet
N(Variablenliste) - anstelle von Summen wird die Anzahl gültiger Werte berechnet

4.) Die Variablen der Variablenliste können durch Komma vom Leerzeichen getrennt werden. Anstelle von

```
SUM (KERNF VERW INW GESIN)
```

kann also geschrieben werden

```
SUM (KERNF,VERW,INW,GESIN)
```

2.2.1 Individuen und Aggregate als Klassifikationsobjekte

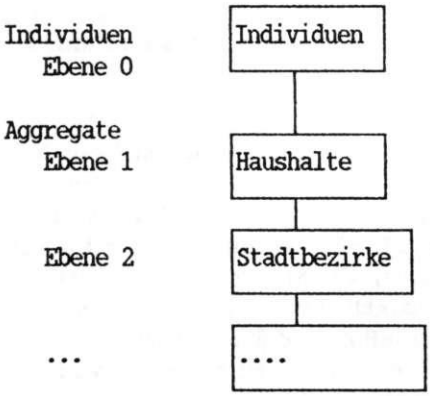
Klassifikationsobjekte können durch Individuen oder Aggregate gebildet werden. Die Aggregate selbst können sich auf unterschiedliche Ebenen beziehen (siehe Abbildung 2.2-2).

Für die Clusteranalyse ist aber nicht die Differenzierung in unterschiedliche Ebenen bedeutsam, sondern die Frage, ob Aggregate oder Individuen feste Zahlenwerte oder Verteilungen in den Klassifikationsmerkmalen besitzen. Die Haushalte in unserem Beispiel gehören der ersten Gruppe an. Sie besitzen nur eine einzige Ausprägung in den Klassifikationsmerkmalen AKERNF, AVERW, AINW und AGESIN. Die aus den Haushalten gebildeten Cluster werden dagegen der zweiten Gruppe angehören.

Zur Messung der Unähnlichkeit zwischen Aggregaten oder Individuen mit Verteilungen in den Klassifikationsmerkmalen können charktcristische Werte der Verteilungen, wie z.B. Modalwerte, Mediane oder Mittelwerte, oder die gesamte Verteilung verwendet werden (s. dazu Abschnitt 2.4.5).

Abbildung 2.2-2:

Individuen und Aggregate unterschiedlicher Ebene



Übungsaufgabe 2:
Gegeben ist folgende Datenmatrix:
Einheiten = Personen
Variablen = BERUF mit den Ausprägungen

1 = selbständig
2 = Unternehmer, Fabrikant, Grundbesitzer
3 = gehobener Beamter
4 = mittlerer Beamter
5 = einfacher Beamter
6 = gehobener Angestellter
7 = mittlerer Angestellter
8 = einfacher Angestellter
9 = Meister
10 = Vorarbeiter
11 = gelernter Arbeiter
12 = angelernter Arbeiter
13 = ungelernter Arbeiter, Taglöhner
14 = sonstiges

EINK (Einkommen) gemessen in Kronen
WFL (verfügbare Wohnfläche) gemessen in Quadratmetern

Diese Daten sollen auf der Datei VZ.DAT stehen. Gesucht ist folgende Klassifikationsdatenmatrix:
Klassifikationsobjekte = Berufe
Klassifikationsmerkmale =
- Einkommensverteilung mit den Ausprägungen 1 = »0 - 1000 Kr«, 2 = »1001 - 2000 Kr«, 3 = »2001 - 5000 Kr« und 5 = « über 5000 Kr«

- durchschnittliche Wohnfläche

Schreiben Sie die Struktur der Ausgangsdatenmatrix, der Klassifikationsdatenmatrix und das entsprechende SPSS·X Programm an. Die aggregierten Daten sollen extern auf der Datei AVZ.DAT gespeichert werden.

23 Vergleichbarkeit und Transformation von Klassifikationsmerkmalen: Eine erste Übersicht

Formal müssen die Klassifikationsmerkmale »vergleichbar« sein. **Vergleichbarkeit** (Kommensurabilität) von Klassifikationsmerkmalen liegt dann vor, wenn die Klassifikationsmerkmale in **derselben Maßeinheit** gemessen werden. Diese Annahme ist z.B. nicht erfüllt, wenn zur Klassifikation der wirtschaftlichen Entwicklung von Staaten die Klassifikationsmerkmale »Pro-Kopf-Bruttosozialprodukt« und »jährliches Wirtschaftswachstum« verwendet werden, da das »Pro-Kopf-Bruttosozialprodukt« in einer bestimmten Währungseinheit als Maßeinheit gemessen wird, das »jährliche Wirtschaftswachstum« dagegen in Prozenten als Maßeinheit. In diesem Fall würden die Ergebnisse ausschließlich durch das Bruttosozialprodukt bestimmt werden, da beim jährlichen Wirtschaftswachstum maximal eine Differenz von 30% empirisch auftreten wird, während beim Bruttosozialprodukt ein Unterschied von 30 Währungseinheiten vernachlässigt werden kann und Unterschiede von 10000 oder mehr Wohnungseinheiten auftreten werden. Aber selbst wenn alle Klassifikationsmerkmale dieselbe Maßeinheit besitzen, wie z.B. der »Anteil des industriellen Sektors am Bruttoinlandsprodukt« und das »jährliche Wirtschaftswachstum«, so ist es doch fraglich, ob diese beiden Klassifikationsmerkmale dieselbe »Maßeinheit« besitzen, ob also eine Differenz von 5% bei der jährlichen Wirtschaftswachstumsrate dasselbe bedeutet wie bei der Industrialisierungsquote, da empirisch die Industrialisierungsquote stärker variieren wird als das jährliche Wirtschaftswachstum (vgl. dazu auch Fox 1982:132).

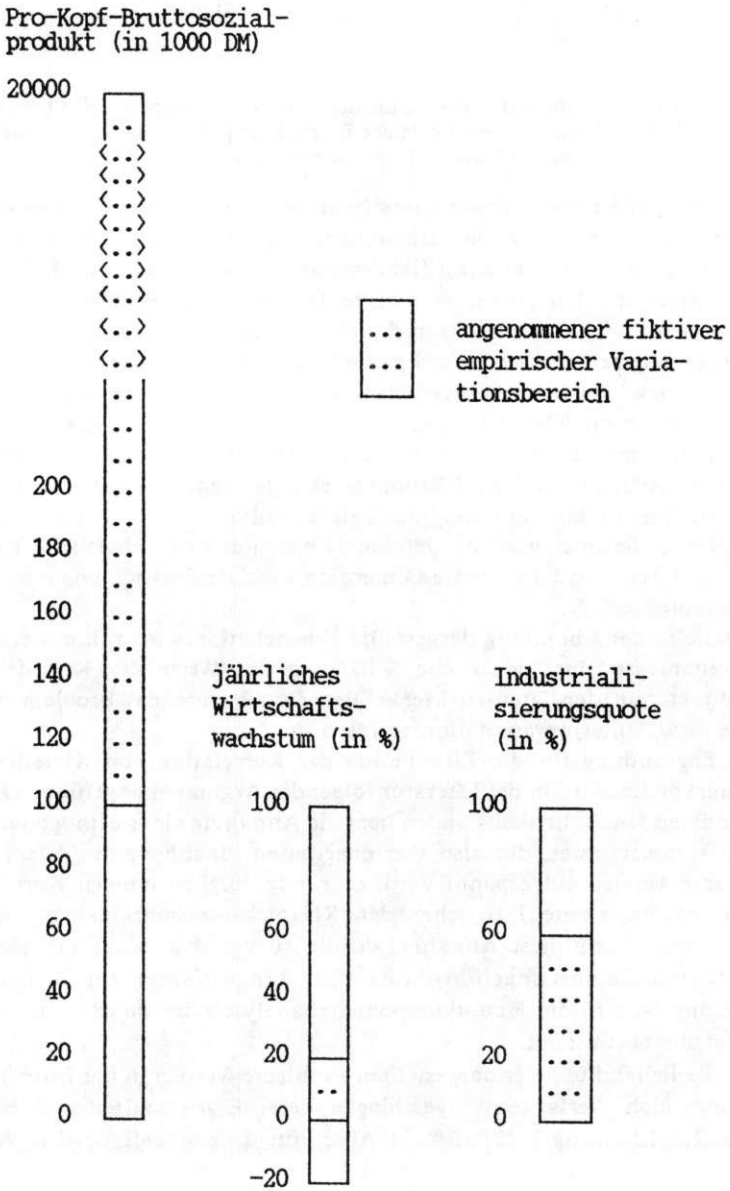
In der Abbildung 2.3-1 ist diese Situation in einem fiktiven Beispiel dargestellt. Man sieht, daß die maximale Differenz im jährlichen Wirtschaftswachstum von 20% beim Bruttosozialprodukt nur ein sehr kleines Intervall auf der Skala einnimmt und bei der Skala der Industrialisierungsquote ebenfalls nur ein Drittel.

Wenn die Klassifikationsmerkmale **unterschiedliches Meßniveau** besitzen, liegt auf jeden Fall **keine Vergleichbarkeit** vor.

Als weitere Ursachen für die Nichtvergleichbarkeit werden in der Literatur genannt (s. dazu die zusammenfassenden Ausführungen in Schlosser 1976: 60-88, Sodeur 1974: 44-59 oder Vogel 1971: 50-78):

Abbildung 2.3-1:

IMerschiedliche Maßeinheiten von Klassifikationsmerkmalen



- **hierarchische Klassifikationsmerkmale,**
- **Über - bzw. Unterrepräsentativität** von Klassifikationsmerkmalen und
- **Korrelation** von Klassifikationsmerkmalen

In der Literatur verwendete synonyme Ausdrücke für diese Probleme sind:

Für hierarchische Klassifikationsmerkmale: bedingte - , komplexe oder primäre und sekundäre Merkmale; und ungleiche Blockbildung der Klassifikationsmerkmale für das Problem der Unter - bzw. Uberrepräsentation.

Hierarchische oder bedingte **Klassifikationsmerkmale** liegen dann vor, wenn das Auftreten eines Klassifikationsmerkmals von dem Auftreten eines oder mehrerer anderer Klassifikationsmerkmale abhängt. So z.B. kann das Auftreten des Klassifikationsmerkmals »derzeitiger Beruf« von dem vorausgehenden Klassifikationsmerkmal »Berufstätigkeit« mit den Ausprägungen »derzeit nicht berufstätig« und »derzeit berufstätig« abhängen.

Über - bzw. Unterrepräsentativität der Klassifikationsmerkmale tritt dann auf, wenn die Klassifikationsmerkmale latente (nicht beobachtbare) Dimensionen messen und dabei jede latente Dimension durch eine unterschiedliche Anzahl von Klassifikationsmerkmalen repräsentiert ist. Dieser Sachverhalt ist in der Abbildung 2.3-2 dargestellt.

In diesem Beispiel wird die latente Dimension »wirtschaftliche Entwicklung durch 2 und die latente Dimension »soziale Entwicklung« durch 3 Indikatoren erfaßt.

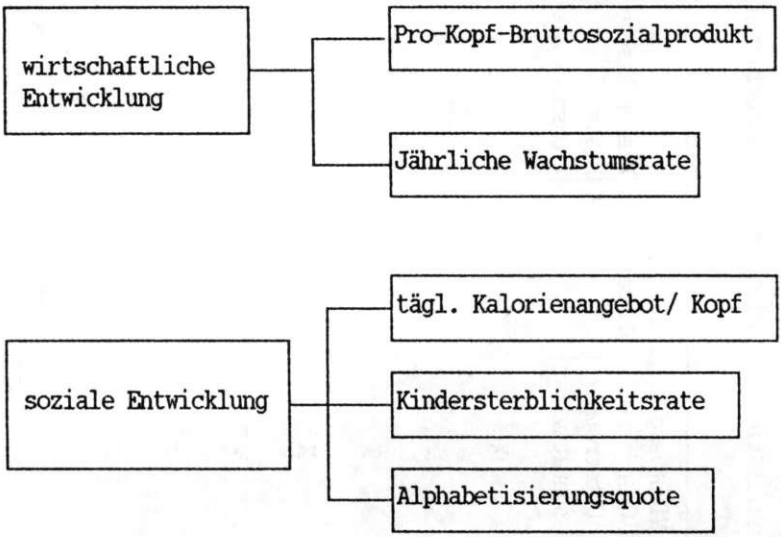
Für die in der Abbildung dargestellte Konstellation wird z.B. mit einer Faktorenanalyse eine empirische Schätzung der Werte der Klassifikationsobjekte auf den latenten Merkmalen gesucht, um das Problem der Über - bzw. Unterrepräsentation zu lösen.

Als Begründung für die **Elimination der Korrelation** von Klassifikationsmerkmalen wird in der Literatur folgendes Argument angeführt: Den verwendeten Unähnlichkeitsmaßen liegt die Annahme eines orthogonalen Klassifikationsraumes, der also von paarweisen, unabhängigen Klassifikationsmerkmalen aufgespannt wird, zugrunde, folglich müssen Korrelationen (Abhängigkeiten) zwischen den Klassifikationsmerkmalen eliminiert werden, damit diese Annahme erfüllt ist (vgl. dazu Kaufman 1985: 470-472 und die dort angeführte Literatur). Ein Verfahren zur Orthogonalisierung ist z.B. die Hauptkomponentenanalyse oder die Berechnung der Mahalanobisdistanz.

Für die Behandlung der dargestellten Probleme werden in der Literatur unterschiedliche Verfahren vorgeschlagen, deren Eigenschaften und technische Durchführung in SPSS-X in Abschnitt 4 behandelt werden. Die Tabelle 2.3-1 gibt einen ersten Überblick über diese Verfahren.

Abbildung 2.3-2:

KLassif ikationsmerkmale als Indikatoren latenter DJnensionen



Ein »x« bedeutet in der Tabelle, daß das entsprechende Verfahren zur Lösung des in der Spalte angeführten Problems geeignet ist. Wichtig für die Anwendung ist, sich das für die Verfahren erforderliche Meßniveau vor Augen zu halten. Wie ist für die einzelnen Verfahren in Tabelle 2.3-2 dargestellt.

Aus der Tabelle 2.3-2 ist ersichtlich, daß eine theoretische und empirische Gewichtung sowie die Anwendung einer Hauptachsentransformation (Hauptkomponentenmethode) oder einer Faktorenanalyse nur durchgeführt werden können, wenn die Klassifikationsmerkmale quantitatives Meßniveau besitzen. Quantitatives Meßniveau muß auch bei der Berechnung der Mahalanobisdistanz vorliegen. Das Informationsmaß von Jardine und Sibson setzt nominales oder ordinales Meßniveau voraus. Keine Voraussetzung bezüglich des Meßniveaus werden bei den verbleibenden Lösungsverfahren getroffen. Abschließend sei aber darauf hingewiesen, daß nominale und ordinale Variablen durch Auflösung in nominale und ordinale Dummies »quasi-quantifiziert« werden können (s. Abschnitt 2.4.4).

Tabelle 2.3-1:
Gründe für die Nichtvergleichbarkeit von Klassifikationsmerkmalen und Lösungsverfahren

Lösungsverfahren:	Gründe für die Nichtvergleichbarkeit: unterschiedliche Maßeinheiten	Über- bzw. Unterrepräsentativität	Korrelation	hierarchische Klassifikationsmerkmale	unterschiedliches Meßniveau
♦ theoretische Gewichtung (Abschnitt 4.2)	X	X	X	.	.
♦ empirische Gewichtung (Abschnitt 4.2)	X	.	.	.	.
♦ Faktorenanalyse (wird nicht behandelt)	.	X	X	.	.
♦ Hauptachsentransformation (Abschnitt 4.2.3)	X	X	X	.	.
♦ Transformation auf nichthierarchische Klassifikationsmerkmale (Abschnitt 4.2.4)	.	.	.	X	.
♦ Reduktion des Meßniveaus	.	.	.	.	X
♦ Berechnung geeigneter Unähnlichkeitsmaße (Abschnitt 4.2.4)	.	.	X (Mahalanobisdistanz)	X (Informationsmaß Jardine & Sibson)	X (Unähnlichkeitsmaß Gower)

Tabelle 2.3-2:

Erforderliches Meßniveau für die einzelnen Verfahren

Lösungsverfahren:	erforderliches Meßniveau:
theoretische Gewichtung	quantitativ
empirische Gewichtung	quantitativ
Faktorenanalyse	quantitativ
Hauptachsentransformation	quantitativ
Transformation auf nicht-hierarchische Klassifikations-merkmale	keine Voraussetzung
Reduktion des Meßniveaus	keine Voraussetzung
Berechnung geeigneter Unähnlichkeitsmaße	
Mahalanobisdistanz	quantitativ
Informationsmaß von Jardine & Sibson	nominal und/oder ordinal
Unähnlichkeitsmaß von Gower	keine Voraussetzung

Übungsaufgabe 3:

a) In einer Untersuchung wurde zur Elimination der Korrelation zwischen den Merkmalen Familienstand und Alter die Mahalanobisdistanz verwendet. Ist dieses Vorgehen zulässig? (Begründen Sie Ihre Antwort)!

b) »Even when all attributes are of the same type, they may be incommensurable. Since commensurability depends upon the possession of a common unit of measurement, only numerical attributes may be strictly commensurable « (Fox 1982: 132)
Bedeutet diese Aussage, daß quantitatives Meßniveau eine notwendige Voraussetzung für Vergleichbaren ist?
Und falls ja. isi diese Aussage richtig?

2A Durchführung einer hierarchischen Clusteranalyse in SPSS-X

In diesem Abschnitt soll anhand unseres Beispiels die Anwendung der SPSS-X Prozedur CLUSTER (SPSS-X 1986: 776-788) dargestellt werden. Die Prozedur CLUSTER ermöglicht

a) eine Berechnung einer Unähnlichkeits - bzw. Ähnlichkeitsmatrix für die Klassifikationsobjekte und
b) die Durchführung einer hierarchisch agglomerativen Clusteranalyse.

2.4.1 Die Grandidee und der Algorithmus hierarchisch agglomerativer Clusteranalyseverfahren

Agglomerative hierarchische Clusteranalyseverfahren sind dadurch gekennzeichnet, daß zu **Beginn der Analyse jedes Klassifikationsobjekt ein selbständiges Cluster** bildet. Diese Cluster werden nun **schrittweise verschmolzen** (agglomeriert), bis nur mehr ein einziges Cluster verbleibt, das alle Klassifikationsobjekte enthält. Cluster, die an einer bestimmten Stelle des Verschmelzungsvorganges zusammengefaßt wurden, können in einem **späteren Schritt** der Verschmelzung **nicht mehr getrennt** werden. Aus diesem Grund werden diese Verfahren hierarchisch bezeichnet. Graphisch läßt sich - wie in Abbildung 2.4-1 gezeigt - der Verschmelzungsvorgang in Form eines **Dendrogramms** darstellen:

In diesem fiktiven Beispiel werden die Klassifikationsobjekte (Cluster) (A) und (B) bei einem Niveau von 2.0 miteinander verschmolzen. Im nächsten Schritt des Verschmelzungsvorganges wird das Cluster (C), das nur durch das Klassifikationsobjekt C gebildet wird, mit dem Cluster (A,B) bei einem Niveau von 3.0 fusioniert. Der darauf folgende Schritt führt zu einer Fusionierung der Cluster (D) und (E) bei einem Niveau von 5.0. Im letzten Schritt des Verschmelzungsvorgangcs werden bei einem Niveau von 10.0 die beiden Cluster (A,B,C) und (D,E) zusammengefaßt.

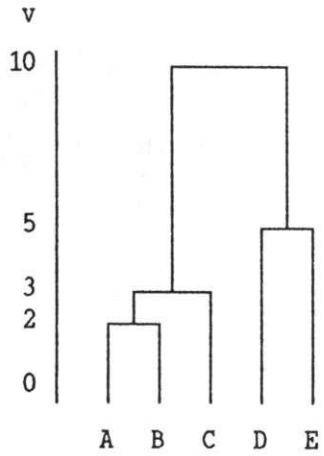
Tabelle 2.4-1:

Verschmelzungsschema für fiktives Beispiel der Abbildung 2.4-1

Schritt	Cluster 1	Cluster 2	Verschmelzungsnivcau
1	A	B	2.0
2	A,B	C	3.0
3	D	E	5.0
4	A,B,C	D,E	10.0

Abbildung 2.4-1:

Darstellung eines Verschmelzungsvorganges
mit Hilfe eines Dendrogranms



Eine andere aber vollkommen äquivalente Darstellung des Verschmelzungsvorganges enthält das sogenannte Verschmelzungsschema (Agglomerationsschema). Für unser fiktives Beispiel hat dieses die in der Tabelle 2.4-1 dargestellte Gestalt.

Die Funktion des Dendrogramms bzw. des Verschmelzungsschemas besteht u.a. darin, die Anzahl homogener Cluster zu bestimmen, in die sich die Klassifikationsobjekte zusammenfassen lassen. Die Anzahl der Cluster wird dabei so bestimmt, daß das Verschmelzungschema von oben nach unten gelesen wird und relativ große Zuwächse notiert werden. Diese großcn Zuwächse bedeuten nämlich, daß an dieser Stelle bereits sehr heterogene Cluster zusammengefaßt werden. Die Tabelle 2.4-2 enthält einige fiktive Ergebnisse von Verschmelzungsprozessen. In der Situation A läßt sich eindeutig zwischen dem 5. und 6. Schritt ein großer Sprung im Verschmelzungsniveau feststellen. In dieser Situation wird man die Anzahl der Cluster auf 5 festsetzen. (Die Zahl 5 erhält man, indem vom letzten Schritt des Verschmelzungsschemas ausgehend rückwärts mit 1, 2, 3, .. gezählt wird).

In der Situation B ist das Bild weniger eindeutig. Man kann die Anzahl der Cluster zwischen den Schritten 6 und 7 und zwischen den Schritten 5

Tabelle 2.4-2:

Die Bestimmung der Anzahl der Cluster aufgrund der im Verschmelzungsschema enthaltenen Information

Situation A (Eindeutige Bestimmung der Zahl der Cluster möglich)		Situation B (Mehrere Lösungen für eine Anzahl der Cluster)		Situation C (Jedes Klassifikationsobjekt bildet ein Cluster)	
Schritt	Niveau	Schritt	Niveau	Schritt	Niveau
1	1.0	1	1.0	1	8.7
2	1.2	2	1.2	2	8.7
3	1.4	3	1.4	3	8.7
4	1.7	4	1.7	4	8.7
5	1.9 *	5	1.9 *	5	8.7
6	8.3	6	4.3	**6**	8.7
7	8.5	7	8.3 *	7	8.7
8	8.6	8	8.5	8	8.7
9	8.7	9	8.7	9	8.7

und 6 festsetzen. Diese Situation wird i.d.R. in der Praxis auftreten. Eine allgemeine Empfehlung für die Entscheidung kann nicht gegeben werden. Es empfiehlt sich vielmehr, die in Frage kommenden Lösungen weiter zu verfolgen und erst in einem späteren Schritt der Analyse eine Entscheidung zu treffen. Allerdings sollte darauf geachtet werden, daß die Anzahl der Cluster nicht zu groß wird, da dadurch die Ergebnisse unüberschaubar werden und eine Interpretation Schwierigkeiten bereitet.

In der Situation C der Tabelle 2.4-2 erkennt man, daß überhaupt kein Zuwachs auftritt, sondern bereits zu Beginn des Verschmelzungsvorganges ein sehr hohes Niveau auftritt. In diesem Fall bilden alle Klassifikationsobjekte ein selbständiges Cluster.

In manchen Analysen, insbesondere bei einer kleinen Anzahl von Klassifikationsobjekten, wie z.B. bei einer Klassifikation von Berufen (s. dazu Kapitel 5) oder von Begriffen in einer Inhaltsanalyse, wird man die durch das Dendrogramm erzeugte Hierarchie ausführlich interpretieren. An eine solche Hierarchie werden bestimmte mathematische Eigenschaf-

ten gestellt, wie z.B. die der **Ultrametrik** (vgl. Bock, 1974: 359-370; Jardine & Sibson, 1971: 49-51) (5). Wir werden hier aber die hierarchischen Verfahren nur als heuristische Verfahren zur Bestimmung »homogener« Klassen verwenden.

Bereits an dieser Stelle muß darauf hingewiesen werden, daß in SPSS·X bei der hierarchischen Clusteranalyse keine Informationen über die Cluster, wie z.B. die durchschnittlichen Merkmalsausprägungen der Cluster in den Klassifikationsmerkmalen, berechnet werden. Die hierarchischen Clusteranalyseverfahren teilen nur mit, in welcher Reihenfolge die Klassifikationsobjekte verschmolzen werden und liefern Anhaltspunkte für die Bestimmung der Anzahl der Cluster. Hat man sich für eine bestimmte Anzahl von Clustern entschieden, kann aus dem Verschmelzungsschema die Zugehörigkeit jedes Klassifikationsobjektes zu einem bestimmten Cluster berechnet werden. Mit dieser Information können die Cluster durch weitere SPSS·X Prozeduren untersucht werden.

Wie wird nun diese Verschmelzung im Detail durchgeführt? Voraussetzung für die Durchführung ist, daß eine Unähnlichkeits - oder Ähnlichkeitsmatrix der Klassifikationsobjekte vorliegt. Für die folgenden Ausführungen soll die in Abbildung 2.4-2 dargestellte Unähnlichkeitsmatrix gegeben sein.

Diese Matrix enthält als Elemente Maßzahlen für die Unähnlichkeit zwischen Klassifikationsobjekten. Diese Unähnlichkeit beträgt beispielsweise 2.0 für die Klassifikationsobjekte A und B, 3.0 für die Klassifikationsobjekte A und C, usw. Ein größerer Zahlenwert bedeutet dabei eine größere Unähnlichkeit. Die **Unähnlichkeitsmatrix** muß eine **symmetrische** Matrix sein: Die Unähnlichkeit zwischen A und B muß gleich der Unähnlichkeit zwischen B und A sein. Empirisch ist diese Annahme beispielsweise nicht erfüllt, wenn die Unähnlichkeit zwischen A und B durch soziometrische Wahlen gemessen wird. Dabei kann A zwar B als ihm sehr ähnlich einstufen, B muß aber keinesfalls dasselbe Urteil abgeben.

Die Unähnlichkeitsmatrix wird oft auch als **Distanzmatrix** bezeichnet. Strenggenommen ist diese Bezeichnung aber nur zulässig, wenn Distanzmaße zur Messung der Unähnlichkeit berechnet werden. Diese erfüllen - im Unterschied zu Unähnlichkeitsmaßen - zusätzlich die sogenannte Dreiecksungleichheit (s. dazu z.B. Kaufmann/Pape 1984: 374-375), was nicht für alle Unähnlichkeitsmaße behauptet werden kann (6). Im folgenden wird deshalb allgemein die Bezeichnung Unähnlichkeitsmatrix gebraucht und von einer Distanzmatrix nur dann gesprochen, wenn für die Unähnlichkeiten Distanzmaße berechnet wurden.

Eine Ähnlichkeitsmatrix enthält zum Unterschied zu einer Unähnlichkeitsmatrix Ähnlichkeitskoeffizienten zwischen den Klassifikationsobjek-

Abbildung 2.4-2:

Fiktive Uhähnlichkeitsmatrix für Rechenbeispiele

	A	B	C	D	E
A	0.0				
B	2.0	0.0			
C	3.0	2.5	0.0		
D	10.0	7.5	6.5	0.0	
E	8.0	9.0	8.5	5.0	0.0

tcn. Ein größerer Zahlenwert bedeutet also größere Ähnlichkeit. Sie muß ebenfalls symmetrisch sein.

Liegt eine Ähnlichkeits - oder eine Unähnlichkeitsmatrix vor, kann der Verschmelzungsprozeß beginnen. Der Algorithmus ist folgender:

Schritt **1:** Jedes Klassifikationsobjekt bildet ein selbständiges Cluster.
Schritt **2:** Suche jene beiden Cluster mit der geringsten Unähnlichkeit (größten Ähnlichkeit) und verschmelze diese zu einem neuen Cluster.
Schritt 3: Berechne neue Unähnlichkeiten (Ähnlichkeiten) zwischen diesem neuen Cluster und den verbleibenden Clustern.
Schritt 4: Überprüfe, ob nur mehr ein Cluster vorliegt. Wenn ja, beende den Verschmelzungsvorgang, wenn nein fahre mit Schritt 2 fort.

Alle hierarchisch agglomerativen Verfahren unterscheiden sich durch die Art und Weise der Berechnung der neuen Un- oder Ähnlichkeiten (vgl. dazu z.B. Sokal *Sc* Sneath 1973: 218-219; Kaufmann *Sc* Pape 1984: 393-394).

Der Algorithmus soll im folgenden nur für den Complete - und Single-Linkage im Detail beschrieben werden.

Complete-Linkage (Methode des weitest entfernten Nachbarns): Der Complete-Linkage versucht, die Unähnlichkeit innerhalb jedes Clusters zu minimieren. Werden zwei Cluster miteinander verschmolzen, dann sollen sich auch die beiden unähnlichsten Klassifikationsobjekte dieses neuen Clusters sehr ähnlich sein. Folglich wird die maximale Unähnlichkeit zwischen dem neuen Cluster und den verbleibenden Clustern als

neues Unähnlichkeitsmaß verwendet. Dadurch ist gewährleistet, daß in den folgenden Verschmelzungsschritten nur bei geringer Unähnlichkeit das neue Cluster mit den verbleibenden Clustern verschmolzen wird. Der Complete-Linkage wird deshalb auch als Methode des weitest entfernten Nachbarns bezeichnet. Diese Analogie soll zur Illustration des zentralen Gedankens des Complete-Linkage dienen:

Abbildung 2.4-3:

Berechnung der neuen Unähnlichkeiten beim Complete-Linkage

maximale Entfernung = neue Unähnlichkeit zwischen beiden Clustern.

Die neue Unähnlichkeit zwischen den beiden Clustern ist identisch mit der Entfernung der weitest entfernten Klassifikationsobjekte Q und Y.

Anwendung des Algorithmus für unsere fiktive Unähnlichkeitsmatrix der Abbildung 2.4-2 führt zu folgenden Ergebnissen:

1. Durchlauf der Schritte 2 bis 4:

In der Unähnlichkeitsmatrix erkennt man, daß die Klassifikationsobjekte (Cluster) A und B die geringste Unähnlichkeit besitzen. Sie werden deshalb im ersten Durchlauf zu einem neuen Cluster zusammengefaßt. Die Unähnlichkeit zwischen dem Cluster (A) und dem verbleibenden Cluster (C) beträgt nun 3.0, die zwischen dem Cluster (B) und dem Cluster (C) dagegen nur 2.5. Beim Complete Linkage wird nun die größte Unähnlichkeit als neues Maß für die Unähnlichkeit zwischen dem Cluster (AB) und dem Cluster (C) verwendet, also der Wert 3.0. Ähnliche Überlegungen für die verbleibenden Cluster (D) und (E) ergeben folgende neue Unähnlichkeitsmatrix (vgl. Abbildung 2.4-4).

2. Durchlauf der Schritte 2 bis 4:

Im nächsten Durchlauf werden zunächst die beiden ähnlichsten Cluster (AB) und (C) verschmolzen und nach derselben Logik - wie im ersten Durchlauf - die neuen Unähnlichkeiten berechnet.

Die Schritte 2 bis 4 werden solange durchlaufen, bis ein einziges Cluster existiert. Als Ergebnis erhält man das in Abbildung 2.4-1 dargestellte Dendrogramm oder das in der Tabelle 2.4-1 wiedergegebene Verschmelzungsschema.

Abbildung 2.4-4:

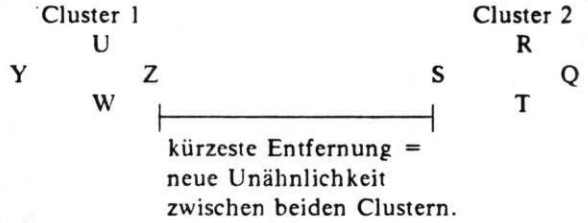
Unähnlichkeitsmatrix nach dem ersten Verschmelzungsschritt beim Complete-Linkage

	AB	C	D	E
AB	0.0			
C	3.0	0.0		
D	10.0	6.5	0.0	
E	9.0	8.5	5.0	0.0

Single-Linkage (Methode des nächsten Nachbarns): Das Verfahren wird auch als **Methode des nächsten Nachbarns** bezeichnet (s. Abbildung 2.4-5):

Abbildung 2.4-5:

Berechnung der neuen Unähnlichkeiten beim Single-Linkage



Die neue Unähnlichkeit zwischen zwei Clustern wird durch die kürzeste Entfernung, die zwischen zwei Klassifikationsobjekten der beiden Cluster besteht, gebildet. In unserem fiktiven Beispiel ergibt der Algorithmus folgende Ergebnisse:

Zunächst werden die beiden ähnlichsten Cluster (A) und (B) verschmolzen. Die neuen Unähnlichkeiten zwischen diesem neuen Cluster (AB) und den verbleibenden Clustern (C), (D) und (E) sind: 2.5 zwischen (AB) und (C), da 2.5 (Unähnlichkeit zwischen (A) und (C)) kleiner ist als 3.0 (Unähnlichkeit zwischen (B) und (C)), 7.5 zwischen (AB) und (D), da 7.5 (Unähnlichkeit zwischen (B) und (D)) kleiner ist als 10.0 (Unähnlichkeit zwischen (A) und (D)).
Im nächsten Durchlauf der Schritte 2 bis 4 werden die beiden ähnlichsten Cluster (AB) und (C) fusioniert. Die neuen Unähnlichkeiten sind: 8.0 zwischen (ABC) und (E) und 6.5 zwischen (ABC) und (D). Als Endergebnis erhält man:

Tabelle 2.4-3:

Verschmelzungsschema für fiktives Beispiel bei Durchführung des Single-Linkage

Schritt	Cluster 1	Cluster 2	Verschmelzungsniveau
1	A	B	2.0
2	A,B	C	2.5
3	D	E	5.0
4	A,B,C	D,E	6.5

Das Niveau von 2.5 in diesem Schema gibt beispielsweise die minimale Unähnlichkeit zwischen den Clustern (AB) und (C) an.

Complete - und **Single*Linkage** stellen zwei **Extreme** in der Berechnung der neuen Unähnlichkeiten dar. Man hat deshalb versucht, **Zwischenlösungen,** die sogenannten **Average - und Centroid·Verfahren,** zu entwickeln (vgl. dazu z.B. Anderberg 1973: 137-145; Sokal & Sneath 1973: 214-245). In der nachfolgenden Tabelle ist die Berechnung der neuen Unähnlichkeiten für diese Verfahren dargestellt, die auch in SPSS·X zur Verfügung stehen.

Neben diesen Verfahren wird in der Literatur häufig das **Verfahren von Ward** als hierarchisch agglomeratives Verfahren angewendet. Dieses Verfahren geht von folgender Überlegung aus: Die Streuung innerhalb der Cluster ist bei quantitativen Klassifikationsmerkmalen ein Maß für die Heterogenität der Cluster. Die Cluster werden nun so gebildet, daß diese Heterogenität minimiert wird. Das Ward·Verfahren bietet darüber hinaus den Vorteil, daß ein Signifikanztest für die Bestimmung der Clusteranzahl durchgeführt werden kann (Kaufman 1985). Der Nachteil dieses Verfahrens besteht darin, daß die Anwendung nur für quantitative Klassifikationsmerkmale zulässig ist und die Unähnlichkeiten durch quadrierte euklidische Distanzen gemessen werden müssen. Eine Anwendung der Logik des Ward-Verfahrens auf nominale Klassifikationsmerkmale stellt die Informationsanalyse (s. z.B. Vogel 1975: 109-129 und 252-291) dar. Sic steht in SPSS·X nicht zur Verfügung.

Angesichts der Vielzahl hierarchisch agglomerativer Verfahren stellt sich unmittelbar die Frage, welches oder welche Verfahren denn nun für eine Clusteranalyse verwendet werden sollen. Eine allgemeine Antwort auf diese Frage kann nicht gegeben werden, vielmehr wird später gezeigt, wie gerade diese Vielfalt zur Überprüfung der Stabilität der erzielten Ergebnisse genutzt werden kann. Formal wird die Anzahl der möglichen Verfahren durch deren Eigenschaften und Anwendungsvoraussetzungen eingeschränkt. Diese sind in der Tabelle 2.4-5 dargestellt.

Tabelle 2.4-4:

Berechnung neuer Unähnlichkeiten bzw. Ähnlichkeiten bei den Average - und Centroidverfahren

Verfahren:	neue Un- bzw. Ähnlichkeiten:
BAVERAGE (average between groups; unweighted pair group method using arithmetic averages (UPGMA))	= Durchschnitt (arithm. Mittel) der Unähnlichkeiten (Ähnlichkeiten) zwischen den Klassifikationsobjekten aus dem Cluster i und j
WAVERAGE (average within groups; weighted pair group method using arithmetic averages (WPGMA)	= Durchschnitt (arithm. Mittel) der Unähnlichkeiten (Ähnlichkeiten) zwischen den Klassifikationsobjekten aus dem Cluster i und j und der Unähnlichkeiten der Klassifikationsobjekte innerhalb von i und j
CENTROID (unweighted centroid)	= Unähnlichkeit (Ähnlichkeit) zwischen den Mittelwerten des Clusters i und j in den Klassifikationsmerkmalen
MEDIAN (weighted centroid)	= Unähnlichkeit (Ähnlichkeit) zwischen den Mittelwerten des Clusters i und j in den Klassifikationsmerkmalen unter Annahme, daß alle Cluster gleich groß sind.

Beim **Ward-Verfahren** können nur **quadrierte euklidische Distanzen** als Unähnlichkeitsmaße verwendet werden. Dadurch werden implizit größere Unterschiede in den Merkmalsausprägungen zweier Klassifikationsobjekte stärker gewichtet als kleine Unterschiede. Ist diese Gewichtung unerwünscht, sollte auf eine Anwendung des Ward-Verfahrens verzichtet wer-

Tabelle 2.4-5:

Anwendungsvoraussetzungen und Eigenschaften agglomerativ hierarchischer Verfahren

Verfahren:	Anwendungsvoraussetzungen:	Eigenschaften:
COMPLETE	nur Ordnungsinformation der Unähnl. geht in die Analyse ein	I.d.R. wird ein heterogenes Restcluster gebildet.
SINGLE	wie beim COMPLETE	Kann zu einer Verkettung führen.
BAVERAGE	Größenunterschiede in den Unähnl. gehen in die Analyse ein	Die Größe und Streuung der Cluster geht in die Analyse ein, deshalb können weit entfernte Cluster mit geringer Streuung verschmolzen werden.
WAVERAGE	wie beim BAVERAGE	
CENTROID	quantitative Klassifikationsmerkmale und quadrierte euklidische Distanz	Inversion des Dendrogramms möglich (Cluster, die zu einem späteren Zeitpunkt verschmolzen werden, können ähnlicher sein als Cluster, die zu einem früheren Zeitpunkt verschmolzen werden implizite Gewichtung durch quadrierte euklidische Distanz
MEDIAN	wie beim CENTROID	
WARD	wie beim CENTROID	implizite Gewichtung durch quadrierte euklidische Distanz

den. Zuder sind für das Verfahren von Ward **quantitative Klassifikationsmerkmale** erforderlich. Diese beiden Einschränkungen gelten auch für die Zentroid-Verfahren (CENTROID und MEDIAN), da nur in diesem Fall die neuen Unähnlichkeiten wie in Tabelle 2.4-4 interpretiert werden können (vgl. Anderberg 1973: 141). Beim **Complete** - und **Single-Linkage** wird nur die **ordinale Information** der Unähnlichkeitsmatrix verwendet. Die Werte der Unähnlichkeitsmatrix können also z.B. quadriert, logarithmiert werden oder es kann die Wurzel gebildet werden, ohne daß sich die Ergebnisse ändern. In die anderen Verfahren dagegen geht das Ausmaß der Unähnlichkeit, also um wieviel unähnlicher sich A und B und A und **C** sind, ein. Auf eine Anwendung des Single-Linkage sollte schließlich wegen der Verkettungseigenschaft verzichtet werden, wenn diese unerwünscht ist.

Übungsaufgabe 4:

a) Führen sie für die nachstehende Ähnlichkeitsmatrix eine Clusteranalyse mit Hilfe des Complete - und Single-Linkage durch. Stellen Sie die Ergebnisse in Form eines Den **Urogramms** und eines Verschmelzungsschemas dar!

	A	B	C	D	E
A	0.				
B	1.	0.			
C	2.	2.	0.		
D	4.	10.	1.	0.	
E	12.	5.	4.	4.	0.

b) In einer Untersuchung wurde für die Klassifikationsmerkmale Familienstand, Geschlecht und Alter das Ward-Verfahren angewendet. Ist dieses Vorgehen zulässig?
Begründen Sie Ihre Antwort!

c) Geben Sie ein Beispiel für eine Clusteranalyse, bei der eine Verkettung erwünscht ist!

2.4.2 P- und D-Maße zur Messung der Ähnlichkeit und Unähnlichkeit zwischen den Klassifikationsobjekten

Eine sehr grobe, aber zentrale Einteilung von Ähnlichkeits - und Unähnlichkeitsmaßen ist die in P - (Produktmaße) und in D-Maße (Distanzmaße) (vgl. Schlosser 1976: 92-151). Den P-Maßen liegt die Berechnung

von **Produkten** in den Merkmalsausprägungen der Klassifikationsmerkmale zwischen zwei Klassifikationsobjekten zugrunde, den **D-Maßen** die Berechnung von **Abständen (Distanzen)** in den Merkmalsausprägungen. Ein Beispiel für ein P-Maß ist die Korrelation zwischen zwei Klassifikationsobjekten. Die aus der Schulmathematik bekannte euklidische Distanz ist ein Beispiel für ein D-Maß. Die Formel zur Berechnung dieser beiden Maßzahlen ist:

$$CORR(i,j) = \frac{\sum_{l}(X_{il}-X_{i.})(X_{jl}-X_{j.})}{[\sum_{l}(X_{il}-X_{i.})^2 \sum_{l}(X_{jl}-X_{j.})]^{1/2}}$$

$$EUCLID(i,j) = [\sum_{l}(X_{il}-X_{jl})]^{1/2}$$

wobei

X_{il} = Ausprägung des Klassifikationsobjektes i in dem Klassifikationsmerkmal l (l = 1,...,m)

$X_{i.} = (\Sigma X_{il})/m$ = Mittelwert des Klassifikationsobjektes i in den Klassifiaktionsmerkmalen l

X_{jl} = Ausprägung des Klassifikationsobjekts j in dem Klassifikationsmerkmal l (l = 1,...,m)

$X_{j.} = (\Sigma X_{jl})/m$ = Mittelwert des Klassifikationsobjektes j in den Klassifikationsmerkmalen l

m = Anzahl der Klassifikationsmerkmale

Für die beiden ersten Haushalte (HH) unseres Datensatzes mit den Merkmalsausprägungen:

	AKERNF	AVERW	AINW	AGESIN
HH1 = [	1.0	0.0	0.0	3.0]
HH2 = [	3.0	3.0	1.0	0.0]

erhält man folgende Werte für CORR(1,2) und EUCLID(1,2):

$X_{1.} = (1.0 + 0.0 + 0.0 + 3.0)/4 = 1.0$

$X_{2.} = (3.0 + 3.0 + 1.0 + 0.0)/4 = 1.75$

$$\begin{aligned}\sum_{l}(X_{1l}-X_{1.})(X_{2l}-X_{2.}) &= (3.0 - 1.75)(1.0 - 1.0) + \\ &+ (3.0 - 1.75)(0.0 - 1.0) + \\ &+ (1.0 - 1.75)(0.0 - 1.0) + \\ &+ (3.0 - 1.75)(3.0 - 1.0) = \\ &= -4.0\end{aligned}$$

$$\sum_{l}(X_{1l} - X_{1.})^2 = (3.0 - 1.75)^2 + (3.0 - 1.75)^2 + (1.0 - 1.75)^2 + (0.0 - 1.75)^2 = 6.75$$

$$\sum_{l}(X_{2l} - X_{2.})^2 = (1.0 - 1.0)^2 + (0.0 - 1.0)^2 + (0.0 - 1.0)^2 + (3.0 - 1.0)^2 = 6.0$$

$$CORR(1,2) = \frac{-4.0}{\sqrt{9.0}\ \sqrt{6.0}} = -0.629$$

$$EUCLID(i,j)^2 = \sum_{l}(X_{il} - X_{jl})^2 = (1.0 - 3.0)^2 + (0.0 - 3.0)^2 + (0.0 - 1.0)^2 + (3.0 - 0.0)^2 = 23$$

$$EUCLID(i,j) = [EUCLID(i,j)]^{1/2} = 4.8$$

Die Korrelation CORR(1,2) ist ein **Maß** für die Ähnlichkeit zwischen zwei Klassifikationsobjekten. Sie läßt sich mit 2(1-CORR(1,2)) - 3.258 in ein Unähnlichkeitsmaß überführen. Das durch diese Transformation erhaltene Unähnlichkeitsmaß entspricht der **quadrierten euklidischen Distanz für standardisierte Klassifikationsobjekte** (7).

Entscheidend für das Verständnis und die Anwendung in SPSS-X ist die Kenntnis der Operationen, die in die Berechnung der Korrelation CORR(iJ) eingehen. Diese sind:

1. Standardisierung der Klassifikationsobjekte.

Bezeichnen wir mit $X_{i.}$ den **Mittelwert des Klassifikationsobjektes** i und mit **Sj** die **Standardabweichung** des Klassifikationsobjektes i in den Klassifikationsmerkmalen, dann werden die Merkmalsausprägungen des Klassifikationsobjektes i standardisiert mit:

$$Z_{il} = (X_{il} - X_{i.})/S_i$$

mit

X_{il} = ursprüngliche Ausprägung des Klassifikationsobjektes i beim Klassifikationsmerkmal l

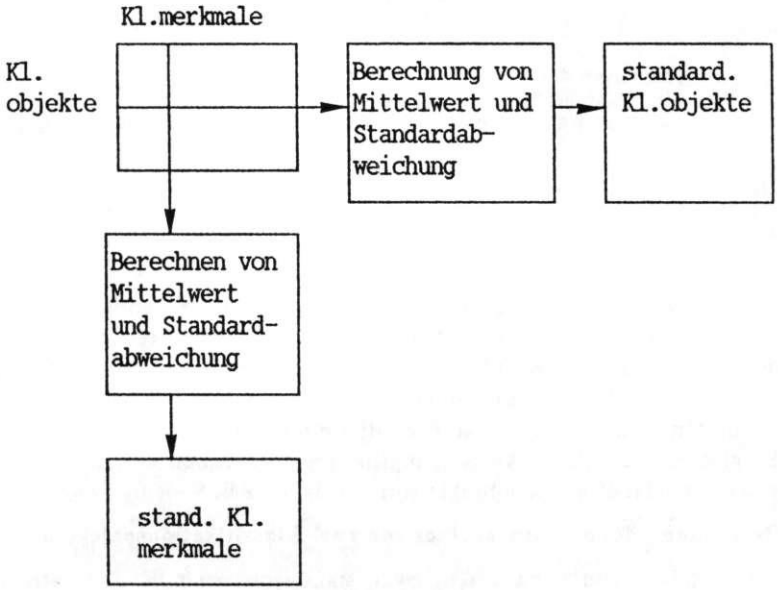
$X_{i.}$ = $(\Sigma X_{il})/m$ = Mittelwert des Klassifikationsobjektes i in den Klassifikationsmerkmalen l (l = 1,2,..,m)

S_i = $\sqrt{(1/m)(\Sigma(X_{il} - X_{i.})^2}$ = Standardabweichung des Klassifikationsobjektes i in den Klassifikationsmerkmalen l (l = 1,2,...,m)

m = Anzahl der Klassifikationsmerkmale

Abbildung 2.4-6:

Standardisierung von Klassifikationsmerkmalen und Klassifikationsobjekten



Diese Standardisierung ist eine **Operation über die Zeilen** der Klassifikationsdatenmatrix, die also für jedes Klassifikationsobjekt durchgeführt wird, während die in den meisten Statistiklehrbüchern bekannte **Z-Transformation** oder Standardisierung eine **Operation über die Spalten** der Klassifikationsdatenmatrix darstellt. Sie wird also für jedes Klassifikationsmerkmal oder allgemein für eine Variable durchgeführt. In der Abbildung 2.4-6 ist der Unterschied zwischen diesen beiden Operationen dargestellt.

Zur Unterscheidung dieser beiden Operationen werden im folgenden die Bezeichnungen **»Standardisierung der Klassifikationsobjekte«** und **»Standardisierung der Klassifikationsmerkmale«** verwendet.

Die Standardisierung der Klassifikationsobjekte bewirkt, daß nur mehr relative Unterschiede zwischen den Klassifikationsobjekten gemessen werden. Die absolute Größe, in unserem Beispiel also die Größe eines

Haushaltes, und die Größe der Unterschiede in den Ausprägungen der Klassifikationsmerkmale spielen keine Rolle mehr. In dem fiktiven Beispiel der Tabelle 2.4-6 würden alle Haushalte nach dieser Standardisierung dieselben standardisierten Merkmalsausprägungen besitzen.

Tabelle 2.4-6:

Konsequenzen einer Standardisierung der Klassifikationsobjekte

	Ausprägungen der Klassifikationsobjekte vor einer Standardisierung				nach einer Standardisierung			
	AKERNF	AVERW	AINW	AGESIN	AKERNF	AVERW	AINW	AGESIN
HH1	1	1	0	0	1	1	-1	-1
HH2	2	2	0	0	1	1	-1	-1
HH3	2	2	1	1	1	1	-1	-1
HH4	3	3	1	1	1	1	-1	-1
HH5	4	4	1	1	1	1	-1	-1

Aus dem Beispiel ist unmittelbar ersichtlich, daß diese Standardisierung von zweifelhaftem Wert ist, da beispielsweise der erste Haushalt mit je einem Mitglied aus der Kernfamilie und aus der Verwandtschaft dieselben standardisierten Merkmalsausprägungen besitzt wie der fünfte Haushalt mit vier Mitgliedern aus der Kernfamilie und einem Verwandten.

Es gibt aber durchaus Anwendungsbeispiele, in denen eine Standardisierung der Klassifikationsobjekte sinnvoll ist (s. z.B. Sodeur, 1974:96).

2. Berechnung des Skalarproduktes von zwei Klassifikationsobjekten.

Das Skalarprodukt zwischen zwei standardisierten Klassifikations-Objekten ist definiert als

$$(\sum_{l} Z_{il}\, Z_{jl})/m$$

Dieses Skalarprodukt ist identisch mit der gesuchten Korrelation CORR(i,j) und entspricht geometrisch dem Cosinus der von den standardisierten Klassifikationsmerkmalen aufgespannten Merkmalsvektoren der beiden Klassifikationsobjekte i und j. In die Messung der Ähnlichkeit durch die Korrelation geht also nur die Richtung dieser beiden Vektoren ein, nicht aber deren Länge.

Der Cosinus kann auch für nicht - standardisierte Klassifikationsobjekte als Ähnlichkeitsmaß berechnet werden. Die Länge der Klassifikationsmerkmalsvektoren spielt dabei ebenfalls keine Rolle.

Technisch kann die Korrelation CORR(i,j) in SPSS X mit diesen beiden Schritten berechnet werden: Zunächst wird die Standardisierung der Klassifikationsobjekte durch COMPUTE-Anweisungen durchgeführt, daran anschließend wird in der Prozedur CLUSTER der COSINUS als Ähnlichkeitsmaß verwendet.

Unähnlichkeitsmaße in SPSS-X.

In der SPSS-X Prozedur CLUSTER selbst kann zwischen folgenden Unähnlichkeitsmaßen gewählt werden:

BLOCK(i,j) = $\Sigma Abs(X_{il} - X_{jl})$ - City-Block oder Manhattenmetrik. Bei der City-Blockmetrik werden die absoluten Abweichungen der Klassifikationsobjekte i und j in jedem Klassifikationsmerkmal berechnet und anschließend aufsummiert. (Summiert wird über alle 1.)

SEUCLID(i,j) = $\Sigma(X_{il} - X_{jl})^2$ = quadrierte Euklidische Distanz. Bei der quadrierten Euklidischen Distanz werden die Abweichungsquadrate von den Merkmalsausprägungen der Klassifikationsobjekte i und j berechnet und anschließend aufsummiert. (Summiert wird über 1.)

EUCLID(i,j) = $[SEUCLID(i,j)]^{1/2}$ - euklidische Distanz.

POWER(i,j/p,r) = $[\Sigma Abs(X_{il} - X_{jl})^p]^{1/r}$ - Minkowskimetrik. Die Minkowskimetrik bietet die Möglichkeit die absoluten Abweichungen der Klassifikationsobjekte i und j in den Klassifikationsmerkmalen unterschiedlich zu gewichten. Je größer der Metrikparameter p gewählt wird, umso stärker werden größere absolute Abweichungen gewichtet. Für p=1 und r=1 erhält man die City-Blockmetrik und für p = 2 und r = 2 die euklidische Distanz.

CHEBYCHEW(iJ) = $\max(Abs(X_{il} - X_{jl}))$ - Distanzmaß von Chebychew. In die Berechnung der Unähnlichkeit geht nur die größte absolute Abweichung von zwei Klassifikationsobjekten in den Klassifikationsmerkmalen ein. (Das Maximum wird also über alle 1 gesucht.)

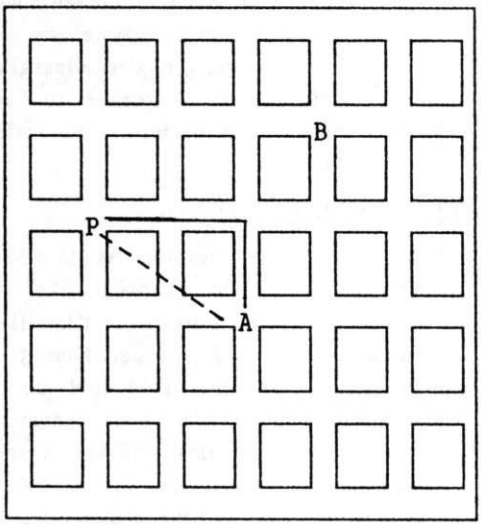
Daneben kann in SPSSX noch der Cosinus COS(i,j)

COSINUS(iJ) - $(\Sigma X_{il}X_{jl})/[\Sigma(X_{il})^2\Sigma(X_{jl})]^{1/2}$

als Unähnlichkeitsmaß berechnet werden (8). Den Unterschied zwischen der City-Blockmetrik, der euklidischen Distanz und dem Cosinus kann folgendermaßen verdeutlicht werden: Gegeben ist eine Ebene, auf der sich zwei Punkte A und B befinden. Die Punkte können durch die direkte Fluglinie oder nur über Kreuzungen, die rechtwinkelig aufeinanderstellen, erreicht werden. Eine Person P befindet sich ebenfalls auf dieser Ebene (vgl. Abbildung 2.4-6a).

Abbildung 2.4-6a:

Graphische Darstellung der City-Blockmetrik und der euklidischen Distanz



— City-Blockentfernung zwischen P und A
.•. euklidische Entfernung zwischen P und A

Die City-Blockmetrik ist in diesem Beispiel die Entfernung zwischen P und A, wenn nur den geraden Straßen und Kreuzungen entlang gegangen werden kann. Der euklidischen Distanz zwischen P und A entspricht die Fluglinienentfernung zwischen P und A. Der Cosinus gibt schließlich an, ob von P aus die beiden Punkte A und B in der gleichen Richtung liegen. Ist der Cosinus 1.0 liegen beide Punkte in derselben Richtung, bei -1.0 in entgegengesetzter Richtung.

Für die beiden ersten Haushalte erhält man die in der Tabelle 2.4-7 dargestelltene Werte.

Die Unähnlichkeitsmaße (BLOCK, SEUCLID, ...) unterscheiden sich durch eine ungleiche Gewichtung der absoluten Abweichungen zweier Klassifikationsobjekte in den Klassifikationsmerkmalen. Nur bei der City-Blockmetrik wird eine Gleichgewichtung aller Abweichungen unabhängig von deren Größen erreicht. Die City-Blockmetrik hat darüber hinaus den Vorteil, daß sie für Klassifikationsmerkmale mit beliebigen Meßniveaus angewendet werden kann.

Tabelle 2.4-7:

Rechenschema für die Berechnung der in SPSS-X verfügbaren Ähnlichkeits- und Unähnlichkeitsmaße

Klassifikationsmerkmale	Klassifikationsobjekte i j	$Abs(X_{il}-X_{jl})$	$(X_{il}-X_{jl})^2$	$X_{il}X_{jl}$	$(X_i,l)^2$	$(X_jl)^2$
AKERNF	1 3	2	4	3	1	9
AVERW	0 3	3	9	0	0	9
AINW	0 1	1	1	0	0	1
AGESIN	3 0	3	9	0	9	0
Σ		9	23	3	10	19
BLOCK(i,j)		9				
SEUCLID(i,j)			23			
EUCLID(i,j)			$\sqrt{23}$			
CHEBYCHEV(i,j)		3				
COSINUS(i,j)					$3/(\sqrt{10})(\sqrt{19})$	

Übungsaufgabe 5:

a) Berechnen Sie für die nachfolgenden Klassifikationsobjekte die City-Blockmetrik, die euklidische Distanz, den Cosinus und die Minkowskimetrik mit p-= 3 und r~ 1.

	XI	X2	X3	X4
KL Objekt 1	1	2	3	2
Kl-objekt 2	10	12	18	16

b) Geben Sie ein Beispiel, in dem eine Standardisierung der Klassifikationsobjekte sinnvoll sein kann.

c) Die Korrelation CORR(ij) zwischen zwei Klassifikationsobjekten i und j läßt sich durch die Transformation

2(l-CORR(i,j))

in ein Unähnlichkeitsmaß transformieren. Man erhält durch diese Transformation die quadrierte euklidische Distanz, wenn die Klassifikationsobjekte standardisiert wurden. Diese Transformation ist auch für den COSINUS (i,j) als Ähnlichkeitsmaß möglich. Erhält man dadurch die quadrierte euklidische Distanz für nicht - standardisierte Klassifikationsobjekte? (Begründen Sie Ihre Antwort!)

2.4.3 Die Standardisierung von Kiassifikationsobjekten

Die Standardisierung der Klassifikationsobjekte zur Berechnung der Korrelation als Ähnlichkeitsmaß muß vor dem Aufruf der SPSS X Prozedur CLUSTER durchgeführt werden. Für unser Beispiel ergibt sich folgen des Programm:

```
TITLE »Standardisierung der Klassifikationsobjekte«
FILE HANDLE AFAMD / NAME « »AFAM.DAT«
GET FILE AFAMD
COMPUTE MKOBJ - MEAN(AKERNF, AVERW, AINW, AGES1N)
COMPUTE SAKORJ - SD(AKERNF, AVERW, AINW, AGESIND)
COMPUTE AKERNF - (AKERNF - MKOBJ)/SAKOBJ
COMPUTE AVERW - (AVERW - MKOBJ)/SAKOBJ
COMPUTE AINW = (AINW - MKOBJ)/SAKOBJ
COMPUTE AGES1N = (AGESIN - MKOBJ)/SAKOBJ
CLUSTER AKERNF AVERW AINW AGESIN
  /MEASURE - COSINUS
```

Durch die FILE HANDLE - Anweisung wird der Name AFAMD der SPSS·X Arbeitsdatei definiert. Diese Datei ist extern unter dem Namen AFAM.DAT abgespeichert. Bei diesen Daten handelt es sich um die auf die Haushalte aggregierten Familienstrukturdaten (s. Hinweis 1 in Abschnitt 2.2). Durch die GET FILE - Anweisung wird die Arbeitsdatei AFAMD »geladen«.

Der Mittelwert der Klassifikationsobjekte in den Klassifikationsmerkmalen wird durch die Anweisung COMPUTE MKOBJ = MEAN (AKERNF, AVERW, AINW, AGESIN) berechnet, die Standardabweichung durch die Anweisung COMPUTE SAKOBJ = SD(AKERNF, AVERW, AINW, AGESIN).

In den folgenden COMPUTE-Anweisungen wird die Standardisierung der Klassifikationsobjekte durchgeführt. Die Anweisung COMPUTE AKERNF= (AKERNF - MKOBJ) / SAKOBJ bewirkt, daß für jedes Klassifikationsobjekt seine standardisierte Ausprägung in dem Klassifikationsmerkmal AKERNF berechnet wird. Die Anweisung COMPUTE AVERW - (AVERW - MKOBJ) / SAKOBJ bewirkt die Berechnung der standardisierten Ausprägungen in dem Klassifikationsmerkmal AVERW für jedes Klassifikationsobjekt, usw..

Die SPSS-X Prozedur CLUSTER wird durch den Befehl CLUSTER aufgerufen. Unmittelbar an den Aufruf anschließend werden die Klassifikationsmerkmale definiert. Die Anweisung MEASURE = COSINUS bedeutet, daß der Cosinus zwischen zwei Klassifikationsobjekten als Ähnlichkeitsmaß berechnet werden soll. Da die Klassifikationsobjekte standardisiert wurden, ist der Cosinus identisch mit der gesuchten Korrelation.

Soll im **Unterschied zur Standardisierung** der Klassifikationsobjekte nur die **absolute Größe** der Klassifikationsobjekte **eliminiert werden,** nicht aber die absoluten Unterschiede in den Merkmalsausprägungen eines Klassifikationsobjektes, dann können z.B. anstelle von Absolutzahlen Anteilswerte in die Analyse einbezogen werden. In unserem Beispiel können diese Anteilswerte wie folgt berechnet werden:

```
TITLE »Berechnung von Anteils werten«
FILE HANDLE AFMAD / NAME - »AFAM.DAT«
GET FILE AFAMD
COMPUTE SKOBJ - SUM ( AKERNF, AVERW, AINW, AGESIN)
COMPUTE AKERNF - AKERNF/SKOBJ
COMPUTE AVERW = AVERW/SKOBJ
COMPUTE AINW = AINW/SKOBJ
COMPUTE AGESIN - AGESIN/SKOR1
```

Die Anweisung TITLE, FILE HANDLE und GET FILE wurden bereits hinlänglich dargestellt und brauchen nicht mehr weiter behandelt werden. Die Anweisung COMPUTE SKOBJ = SUM(AKERNF, AVERW, AINW, AGESIN) bewirkt, daß für jeden Haushalt (Klassifikationsobjekt) die Anzahl der im Haushalt lebenden Personen (= Summe der Klassifikationmerkmale AKERNF, AVERW, AINW und AGESIN) berechnet und der Variablen SKOBJ zugewiesen wird. In den nachfolgenden COMPUTE - Anweisungen werden die gesuchten Anteilswerte berechnet.

Übungsaufgabe **6:**

a) Berechnen Sie in der Tabelle 2.4-6 anstelle der standardisierten Klassifikationsobjekte die entsprechenden Anteilswerte und verdeutlichen Sie sich die Unterschiede!

b) In der Tabelle 2.4-6 sollen anstelle von Anteilswerten die Klassifikationsobjekte um ihre Mittelwerte zentriert werden ($Z_{ij} = X_{il} - X_{i.}$).

Führt dieses Vorgehen ebenfalls zu einer Elimination der absoluten Höhe der Klassifikationsobjekte? Erhält man bei der Verwendung der City-Blockmeirik dieselben Unähnlichkeitswerte wie bei der Verwendung von Anteilswerten? (Begründen Sie Ihre Antwort!)

c) Schreiben Sie für die familialen Haushaltsdaten ein SPSS X Programm, das diese Mittelwertzentrierung durchführt.

2.4.4 Die Anwendung der City-Block metrik für Klassifikationsmerkmale beliebigen Meßniveaus

Die Anwendung der bisher behandelten Ähnlichkeits - und Unähnlichkeitsmaße setzt quantitatives Meßniveau der Klassifikationsmerkmale voraus. Darüber hinaus müssen die Klassifikationsobjekte durch einen einzigen Zahlenwert in den Klassifikationsmerkmalen gekennzeichnet sein. In diesem Abschnitt wird das Vorgehen bei ordinalen und nominalen Klassifikationsmerkmalen dargestellt. Die Strategien für den Fall, daß die Klassifikationsobjekte durch eine Verteilung in den Klassifikationsmerkmalen gekennzeichnet sind, werden im nächsten Abschnitt behandelt.

Für ordinale, nominale und insbesondere für binäre Klassifikationsmerkmale wurden in der Literatur eine Vielzahl von Ähnlichkeits - und Unähnlichkeitsmaßen entwickelt (s. dazu z.B. die in der SPSSX Prozedur PROXIMITY verfügbaren Koeffizienten; SPSS Inc. 1986: 732-735).

Für praktische Zwecke ausreichend dürfte aber die Tatsache sein, daß die **City-Block metrik** für **Klassifikationsmerkmale mit beliebigem** Meßniveau angewendet werden kann. Für nominale Klassifikationsmerkmale können darüber hinaus alle D-Maße zur Messung der Unähnlichkeit verwendet werden. Die dafür erforderlichen Operationen sollen im folgenden kurz dargestellt werden.

Gegeben sei ein nominales Klassifikationsmerkmal A mit k Ausprägungen. Wir wissen bereits, daß sich A in k **Dummy-Variablen** Aj auflösen läßt, wobei Aj den Wert 1 annimmt, wenn A die Ausprägung 1 besitzt. Für alle anderen Ausprägungen von A ist A, gleich 0. Für diese Dummy-Variablen kann nun die CityBlockmetrik berechnet werden:

$$CITY(i,j) = \sum_{l} Abs(A_{il} - A_{jl})$$

wobei $A_{.}$ = Ausprägung des Klassifikationsobjektes i in der Dummy-Variablen Aj

A_{jl} = Ausprägung des Klassifikationsobjektes j in der Dummy-Variablen A_l

Die City-Blockmetrik kann für ein nominales Klassifikationsmerkmal nur zwei Werte annehmen. Den Wert 0, wenn die beiden Klassifikationsobjekte i und j dieselbe Ausprägung in A besitzen, und den Wert 2.0 bei unterschiedlichen Ausprägungen in A.

Grundsätzlich kann dieses Vorgehen auch für ordinale Klassifikationsmerkmale gewählt werden. Allerdings geht dabei die ordinale Information verloren. Sei beispielsweise das ordinale Klassifikationsmerkmal C das in Einkommensgruppen gemessene Einkommen (vgl. Tabelle 2.4-8) einer Menge von Personen (Klassifikationsobjekte), dann werden Personen mit einem Einkommen von DM 0 bis 1000 und von DM 1001 bis 2000 als gleich unähnlich bezeichnet wie Personen mit einem Einkommen von DM 0 bis 1000 und von DM 5001 bis 8000.

Dieser Nachteil kann durch das sogenannte »additive coding« (Vogel 1975: 73-77) beseitigt werden. Beim »additive coding« werden ordinale Dummy - Variablen nach folgender Vorschrift gebildet:

$$C_i \quad \begin{cases} = 1 \text{ wenn die Ausprägung von } C \geq i \\ = 0 \text{ sonst} \end{cases}$$

Die Bildung von ordinalen Dummy-Variablen ist in der Tabelle 2.4-8 dargestellt.

Berechnet man für die ordinalen Dummy Variablen die City-Blockmetrik, dann wird die Anzahl der Ausprägungen, die zwischen zwei Klassifikationsobjekten liegt, gezählt. Die City-Blockmetrik beträgt beispielsweise in der Tabelle 2.4-8 zwischen zwei Personen mit einem Einkommen von DM 0 bis 1000 und von DM 1001 bis 2000 gleich 1.0. Dagegen würde sie für zwei Personen mit einem Einkommen von DM 0 bis 1000 und von DM 5001 bis 8000 den Wert 3.0 annehmen.

Die Anwendung anderer D-Maße - als der City-Blockmetrik - würde zu einer Gewichtung der Unterschiede in der Anzahl der Ausprägungen führen und deshalb mehr als die reine ordinale Information verwenden. Ihre Anwendung ist deshalb nur sinnvoll, wenn diese Gewichtung inhaltlich theoretisch begründet werden kann.

Im Unterschied zu nominalen Klassifikationsmerkmalen hängt der maximale Wert, den die City-Blockmetrik bei ordinalen Klassifikationsmerkmalen annehmen kann, von der Anzahl der Ausprägungen ab. Er beträgt k - 1, wobei k die Anzahl der Ausprägungen ist. Werden also mehrere ordinale Klassifikationsmerkmale in die Analyse einbezogen, ist eine Gewichtung mit k oder mit k - 1 sinnvoll, um die Vergleichbarkeit der Klassifikationsmerkmale zu erreichen.

Tabelle 2.4-8:

Auflösung eines ordinalen Klassifikationsmerkmals in nominale und ordinale Dummy-Variablen

Ordinales Klassifikations-merkmal (Einkommen) in DM:	nominale Dummy-Variablen:						ordinale Dummy-Variablen:					
	C_1	C_2	C_3	C_4	C_5	C_6	C_1	C_2	C_3	C_4	C_5	C_6
0 - 1000	1	0	0	0	0	0	1	1	1	1	1	1
1001 - 2000	0	1	0	0	0	0	0	1	1	1	1	1
2001 - 3000	0	0	1	0	0	0	0	0	1	1	1	1
3001 - 5000	0	0	0	1	0	0	0	0	0	1	1	1
5001 - 8000	0	0	0	0	1	0	0	0	0	0	1	1
über 8000	0	0	0	0	0	1	0	0	0	0	0	1

Das nachfolgende Beispiel zeigt, wie diese Gewichtung und die Bildung von ordinalen Dummy-Variablen in SPSSX durchgeführt werden kann. Dabei wird angenommen, daß die SPSSX Arbeitsdatei die beiden ordinalen Klassifikationsmerkmale EINK (Einkommen) mit den 6 Ausprägungen

1 = 0 - 1000 DM
2 = 1001 - 2000 DM
3 = 2001 - 3000 DM
4 = 3001 - 5000 DM
5 = 5001 - 8000 DM
6 = über 8000 DM

und ALTER (Alter) mit den 4 Ausprägungen

1 = 0 - 25 Jahre
2 = 26 - 40 Jahre
3 = 41 - 60 Jahre
4 = über 60 Jahre

enthält. Die **SPSS-X Befehle** zur Bildung und Gewichtung der **ordinalen Dummy-Variablen** sind:

```
COMPUTE DEINK1-0          DO REPEAT DEINK-DEINK1 TO DEINK6
COMPUTE DEINK2-0          COMPUTE DEINK-0
COMPUTE DEINK3-0          END REPEAT
COMPUTE DEINK4-0
COMPUTE DEINK5-0
COMPUTE DEINK6-0
IF (EINK GE 1) DEINK1 - 1 DO REPEAT DEINK-DEINK1 TO DEINK6
IF (EINK GE 2) DEINK2- 1     /WERT« 1 TO 6
IF (EINK GE 3) DEINK3= 1  IF (EINK GE WERT) DEINK- 1./5.
IF (EINK GE 4) DEINK4- 1  END REPEAT
IF (EINK GE 5) DEINK5-1
IF (EINK GE 6) DEINK6- 1
COMPUTE DALT1 -0          DO REPEAT DALT-DALT1 TO DALT4
COMPUTE DALT2 - 0         COMPUTE DALT - 0
COMPUTE DALT3 - 0         END REPEAT COMPUTE DALT4 - 0
IF (ALTER GE 1) DALT1 - 1 DO REPEAT DALT-DALT1 TO DALT4
IF (ALTER GE 2) DALT2- 1     /WERT- 1 TO 4
IF (ALTER GE 3) DALT3- 1  IF (ALTER GE WERT) DALT- 1/3.
IF (ALTER GE 4) DALT4- 1  END REPEAT
COMPUTE DEINK1 - DEINK1 /5.
COMPUTE DEINK2 —DEINK2/5.
COMPUTE DEINK3 —DEINK3/5.
COMPUTE DEINK4-DEINK4/5.
COMPUTE DEINK5 = DEIN K 5/5.
COMPUTE DEINK6—DEINK6/5.
COMPUTE DALT 1 - DALT 1/3.
COMPUTE DALT2-DALT2/3.
COMPUTE DALT3 — DALT3/3.
COMPUTE DALT4 = DALT4/3.
```

Auf der linken Seite dieses Beispieles sind die ausführlichen SPSS·X Befehle dargestellt, die rechte Seite enthält eine äquivalente, aber wesentlich kürzere Schreibweise dieser Operationen. Die ersten 6 COMPUTE - Anweisungen auf der linken Seite bewirken, daß die ordinalen Dummy·Variablen DEINK1, DE1NK2,..,DEINK6 initialisiert und ihre Werte auf 0 gesetzt werden. Diese Operationen können durch die Verwendung einer DO REPEATSchleife wesentlich kürzer gefaßt werden (s. die rechte Seite). Die 6 COMPUTE - Anweisungen auf der linken Seite sind vollkommen identisch mit der DO REPEAT - Schleife:

```
DO REPEAT DEINK = DE1NK1 TO DE1NK6
COMPUTE DEINK = 0
END REPEAT
```

Die DO REPEAT - Anweisung bewirkt, daß eine Schleife definiert wird. Diese Schleife wird von der Variablen DEINK sechsmal mit DEINK 1 beginnend bis DEINK6 durchlaufen. Für jede dieser Variablen wird die COMPUTE - Anweisung ausgeführt.

Durch die IF - Anweisungen werden auf der linken Seite die **ordinalen Dummy** - Variablen gebildet. Die Anweisung IF (EINK GE 1)

DEINK1 = 1 bewirkt, daß alle Klassifikationsobjekte mit einem Einkommen ≥ DM 0 in der ordinalen Dummy-Varibalen DEINK1 den Wert 1 erhalten. Die Anweisung IF (EINK GE 2) DEINK2 • 1 führt dazu, daß alle Klassifikationsobjekte mit einem Einkommen ≥ DM 1001 in der ordinalen Dummy-Variablen DEINK2 den Wert 1 erhalten, usw.. Auch diese Befehle kann man durch die Verwendung einer DO REPEAT - Schleife kürzer ausführen. Durch die Anweisung DO REPEAT DEINK = DEINK1 TO DEINK2/ WERT = 1 TO 6 wird eine Schleife definiert, die mit der Variablen DEINK und parallel dazu mit der Variablen WERT durchlaufen wird. Im ersten Schleifendurchlauf haben die Variablen DEINK und WERT die Ausprägungen DEINK1 und 1, im zweiten Schleifendurchlauf die Ausprägung DEINK2 und 2 usw.. In der Schleife werden die Dummy-Variablen DEINK1, DEINK2,.. unmittelbar gewichtet, anstelle eines Wertes von 1.0. Wenn die IF-Bedingung zutrifft, erhalten sie den Wert *US.* (5= die Anzahl der Ausprägungen minus 1). In der rechten Spalte wird diese Gewichtung durch COMPUTE-Anweisungen am Ende des Programmes ausgeführt. Durch die Anweisung COMPUTE DEINK1 = DEINK1/5. wird die ordinale Dummy-Variable DEINK1 gewichtet, durch die Anweisung COMPUTE DEINK2 = DEINK2/5. die ordinale Dummy-Variable DEINK2, usw..

Für das ordinale Klassifikationsmerkmal ALTER werden analoge Operationen durchgeführt.

Übungsaufgabe 7: Gegeben ist die SPSS-X Datei mit Personen als Klassifikationsobjekte und den Klassifikationsmerkmalen:

GRUND (Grundbesitz) mit den Ausprägungen:

1 = 0 bis unter 1 ha
2 = 1 bis unter 2 ha
3 = 2 bis unter 5 ha
4 = über 5 ha

REL (Religionszugehörigkeit) mit den Ausprägungen:

1 = röm.Kath.
2 = evangelisch
3 = jüdisch
9 = sonstiges

HERK (Herkunft) mit den Ausprägungen:

1 = einheimisch
2 = Naheinwanderer
3 = Femeinwanderer

Die Datei ist extern unter dem Namen P.DAT abgespeichert.
Schreiben Sie ein SPSS-X Programm, in dem die Klassifikationsmerkmale entsprechend ihrem Meßniveau in ordinale und nominale Dummy Variablen aufgelöst und gewichtet werden.

2.4.5 P- und D-Maße für Individuen und Aggregate

Formal können die Klassifikationsobjekte in zwei Arten eingeteilt werden (vgl. Abschnitt 2.2.1):

- Klassifikationsobjekte mit einer festen Ausprägung in den Klassifikationsmerkmalen und
- Klassifikationsobjekte mit einer Verteilung in den Klassifikationsmerkmalen.

Beide Gruppen von Klassifikationsmerkmalen können Individuen oder Aggregate sein. Für die erste Gruppe lassen sich problemlos die dargestellten P - und D-Maße anwenden. Für die zweite Gruppe können zwei Strategien verfolgt werden.

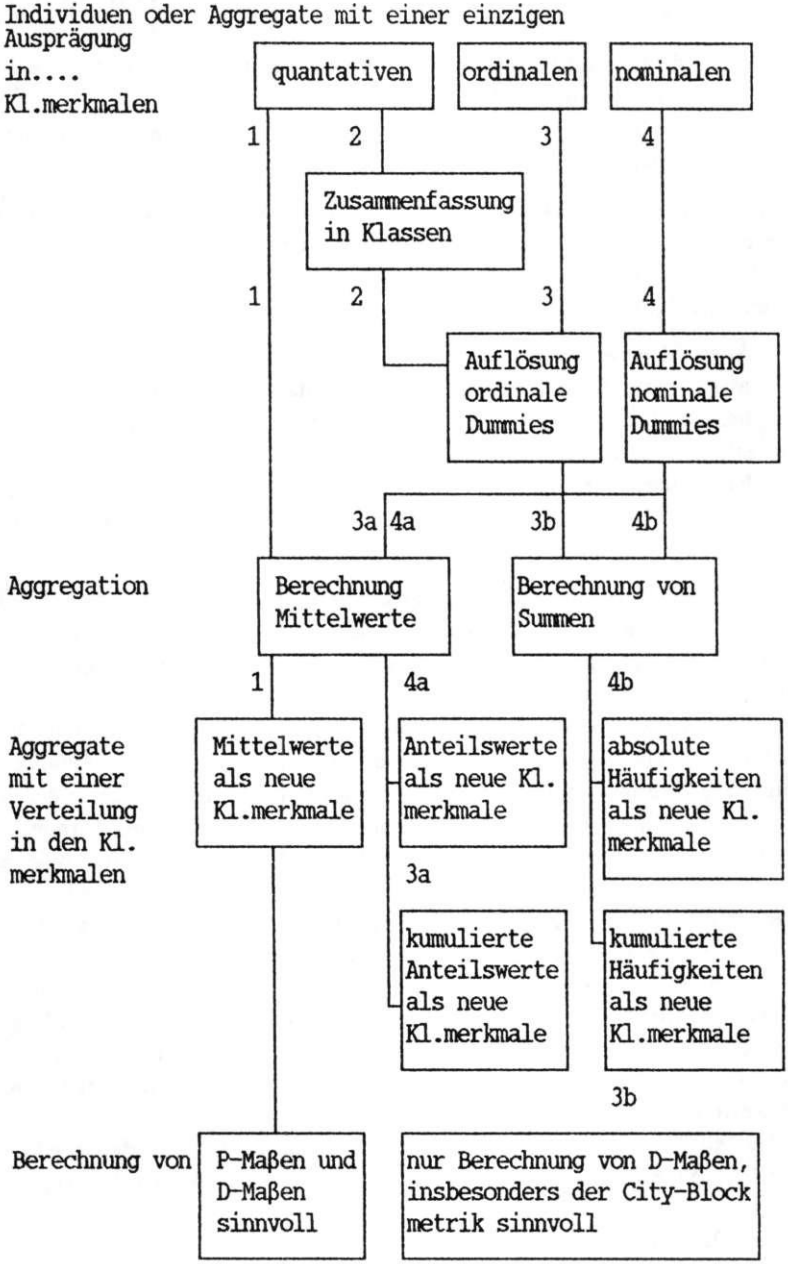
- Auswahl eines charakteristischen Merkmals der Verteilung, wie z.B. Mittelwert, Median, Perzentile oder Modalwert. Durch dieses Vorgehen werden die Klassifikationsobjekte der zweiten Gruppe auf Klassifikationsobjekte der ersten Gruppe reduziert.
- Einbeziehen der Verteilung in die Berechnung der Ähnlich - oder Unähnlichkeitsmaße. Für diese Strategie ist eine Auflösung in Dummy-Variablen erforderlich.

Die Anwendung dieser beiden Strategien ist in der Abbildung 2.4-7 dargestellt. Ausgangspunkt dabei ist, daß eine Menge von Klassifikationsobjekten aggregiert wird und die so entstandenen Aggregate eine Verteilung in den Klassifikationsmerkmalen besitzen.

Bei quantitativen Klassifikationsmerkmalen können unmittelbar die Mittelwerte als neues Klassifikationsmerkmal verwendet werden (Pfad 1). Eine Berechnung von D - und P-Maßen ist möglich. Auf der anderen Seite können quantitative Klassifikationsmerkmale in ordinale Dummies aufgelöst werden. Dafür wird in vielen Fällen eine Zusammenfassung in Klassen erforderlich sein (Pfad 3). Für die aus quantitativen, ordinalen oder nominalen Klassifikationsmerkmalen erzeugten Dummy-Variablen können bei der Aggregation Mittelwerte oder Summen berechnet werden. In Abhängigkeit vom Vorgehen erhält man für nominale Klassifikationsmerkmale Anteilswerte bei der Berechnung von Mittelwerten (Pfad 4a) und absolute Häufigkeiten bei einer Summation (Pfad 4b). Für ordinale oder quantitative Klassifikationsmerkmale erhält man kumulierte Anteilswerte (Pfad 3a) oder absolute Häufigkeiten (Pfad 3b).

Die Möglichkeit bei ordinalen Klassifikationsmerkmalen den Median oder Perzentile und bei nominalen Klassifikationsmerkmalen den Modalwert als charakteristisches Verteilungsmerkmal auszuwählen ist in der Abbildung 2.4-7 nicht eingetragen.

Abbildung 2.4-7: Berechnung von D- und P-Maße für Klassifikationsobjekte unterschiedlichen ***Typus***



Übungsaufgabe 8: Gegeben sei die Datei P.DAT der Aufgabe 7, die zusätzlich das Klassifikationsmerkmal ALTER gemessen in Jahren enthält. Die Datei wurde über die Variable BERUF wie folgt auf aggregiert:

```
FILE HANDLE PDAT / NAME - »P.DAT«
GET FILE - PDAT
RECODE REL(9«4)
DO REPEAT DREL = DREL1 TO DREL4 / WERT - 1 TO 4
COMPUTE DREL - 0
IF (REL - WERT ) DREL - 1
END REPEAT
DO REPEAT DHERK « DHERK1 TO DHERK3 / WERT - 1 TO 3
COMPUTE DHERK - 0
IF (HERK GE WERT) DHERK - 1/2
END REPEAT
DO REPEAT DGRUND - DGRUND1 TO DGRUND4 / WERT - 1 TO 4
COMPUTE DGRUND - 0
IF (DGRUND GE WERT) DGRUND - 1/3
END REPEAT
AGGREGATE OUTFILE « *
      /AALT - MEAN(ALTER)
      /AGRUND1, AGRUND2, AGRUND3, AGRUND4, AREL1,
      AREL2, AREL3 AREL4 AHERK1 AHERK2 AHERK3 - MEAN
      (DGRUND1, DGRUND2, DGRUND4, DREL1, DREL2, DREL3,
      DREL4, DHERK1, DHERK2, DHERK3)
      /BREAK - BERUF
```

Beschreiben Sie das SPSS X Programm und die erhaltenen neuen Klassifikationsmerkmale ausführlich.

2.5 Die Bestimmung der Anzahl der Cluster und ihre Beschreibung für die familialen Haushaltsdaten

In Abschnitt 2.4.1 wurde abstrakt gezeigt, wie mit Hilfe der agglomerativ hierarchischen Verfahren die Anzahl der Cluster, die einer empirischen Klassifikationsdatenmatrix zugrundeliegen, bestimmt werden kann. Dieses Vorgehen soll nun anhand unserer konkreten empirischen Daten demonstriert werden. Es wurde folgendes SPSS X Programm gerechnet.

```
TITLE »Hierarchische Clusteranalyse der osttiroler Haushalte«
FILE HANDLE AFAMD / NAME = »AFAM.DAT«
GET FILE AFAMD
CLUSTER AKERNF AVERW AINW AGESIN
  /MEASURE «= BLOCK
  /METHOD = COMPLETE
  /PRINT - SCHEDULE
  /PLOT - DENDROGRAM
```

Durch die TITLE-Anweisung erhält das Programm einen Namen. Die SPSS-X Arbeitsdatei wird durch die FILE HANLE - und GET FILE - Befehle definiert und »geladen«. Durch die Anweisung CLUSTER wird die SPSS-X Prozedur CLUSTER aufgerufen. Unmittelbar an diesen Aufruf anschließend müssen die Klassifikationsmerkmale definiert werden, die in die Analyse einbezogen werden sollen, in unserem Beispiel also die Klassifikationsmerkmale AKERNF, AVERW, AINW und AGESIN. Als Unähnlichkeitsmaß wird in diesem Programm die City-Blockmetrik wegen der Anweisung MEASURE = BLOCK verwendet. Die City-Blockmetrik wurde gewählt, da alle Unterschiede in den Ausprägungen der Klassifikationsmerkmale gleich gewichtet werden. Die Anweisung METHOD - COMPLETE bewirkt, daß der Complete-Linkage als hierarchisches Clusteranalyseverfahren verwendet wird. Wir haben uns für dieses Verfahren in einem ersten Schritt der Analyse entschieden, da es zu möglichst homogenen Clustern führt. Durch die Anweisung PRINT = SCHEDULE wird die Ausgabe des Verschmelzungsschemas festgelegt. Das ist die Voreinstellung, die PRINT-Anweisung hätte deshalb entfallen können. Die PLOTAnweisung bewirkt, daß ein Dendrogramm ausgedruckt wird. Die Voreinstellung für die PLOTAnweisung ist die Ausgabe eines sogenannten Eiszapfendiagramms, das i.d.R sehr unübersichtlich ist.

Dieses SPSS-X Programm ergibt die in der Tabelle 2.5-1 zusammengefaßten Ergebnisse.

In die Tabelle wurden die englischen Bezeichnungen, wie sie in SPSS-X verwendet werden, übernommen. In diesem Verschmelzungsschema läßt sich ein Zuwachs von 2 Einheiten zwischen dem 155. und 156. sowie zwischen dem 156. und 157. Schritt erkennen, also bei einem Übergang von 4 zu 3 sowie von 3 zu 2 Clustern. Bis zum 155. Schritt wächst der Unähnlichkeitskoeffizient kontinuierlich um eine Einheit. Bei einem Übergang von 2 zu einem Cluster beträgt der Zuwachs schließlich 5 Einheiten. Allerdings beträgt die maximale Unähnlichkeit bei 4 Clustern bereits 12 Einheiten. Die beiden Cluster, die in diesem Schritt verschmolzen werden, nämlich das Cluster 1, das den ersten und weitere Haushalte enthält, und das Cluster 2, das den zweiten und andere Haushalte enthält, unterscheiden sich also durch 12 Einheiten.

Zu beachten ist, daß in SPSS X nur das erste Klassifikationsobjekt eines Clusters ausgedruckt wird und die Klassifikationsobjekte fortlaufend mit 1 beginnend nummeriert werden. Letzteres erschwert in unserem Beispiel die Identifkation der Klassifikationsobjekte »Haushalte«, die durch die Variablen HNR (Hausnummer) und HHNR (Haushaltsnummer) gekennzeichnet sind. Allerdings ist es möglich, den Klassifikationsobjekten Namen zu geben (s. dazu das Beispiel in Kapitel 3).

Bereits an diesem Punkt der Analyse stellt sich die Frage, ob eine 4 oder 3 Clusterlösung weiter analysiert oder ob nicht für die weitere Analyse

Tabelle 2.5-1:

Verschmelzungsschema der familialen Haushaltstrukturen für Complete-Linkage bei Anwendung der City-Blockmetrik

	Clusters combined		
Stage	Cluster 1	Cluster 2	Coeffizient
•			
•			
140	19	28	5.000
141	6	10	5.000
142	1	93	6.000
143	12	24	6.000
144	8	16	6.000
145	3	5	6.000
146	2	4	6.000
147	12	95	7.000
148	6	58	7.000
149	13	37	7.000
150	19	52	8.000
151	1	3	9.000
152	13	19	10.000
153	2	8	10.000
154	6	12	11.000
155	1	2	12.000
156	13	30	14.000
157	1	6	16.000
158	1	13	21.000

eine Lösung mit mehr Clustern verwendet, für die die maximale Unähnlichkeit einen geringeren Wert annimmt, oder ob nicht mit einer feineren Zusammenfassung gearbeitet werden soll. Zur Bestimmung dieser größeren Clusteranzahl fehlt ein formales Kriterium, da bis zum 155. Schritt des Verschmelzungsniveau (Coeffizient) kontinuierlich um 1 wächst.

Auf der anderen Seite sind uns aber die familialen Strukturen der ersten drei Haushalte bekannt (vgl. Tabelle 2.1-2). Der erste Haushalt setzt sich - zur Erinnerung - aus einem Haushaltsvorstand, dem Pfarrer, und drei Angehörigen des Gesindes zusammen, der zweite dagegen aus drei Mitgliedern der Kernfamilie, drei Mitgliedern der Verwandtschaft und einem

Inwohner. Diese Zusammenfassung ist aber wenig erwünscht, wird aber im 155-ten Schritt vorgenommen. Bereits an dieser Stelle läßt sich sagen, daß diese 4-Clusterlösung bestenfalls eine sehr grobe erste Klassifikation der Haushalte darstellen wird, da die beiden Haushalte 1 und 2 zusammengefaßt werden oder es sich dabei um ein Restcluster handelt (s. Tabelle 2.4-5).

Wir wollen zunächst aber die 4 - bzw. 3-Clusterlösung weiterverfolgen und diese vier Cluster durch ihre Ausprägungen in den Klassifikationsmerkmalen beschreiben. Diese Beschreibung ist innerhalb der SPSSX Prozedur CLUSTER nicht möglich. Das Vorgehen ist deshalb folgendes: In der SPSSX Prozedur CLUSTER wird zunächst die Zugehörigkeit jedes Klassifikationsobjektes zu einem der drei Cluster zwischengespeichert. In einem weiteren Schritt wird die Prozedur BREAKDOWN (SPSSX 1986: 372-383) aufgerufen, in der die Mittelwerte und Standardabweichungen der drei Cluster berechnet werden. Das entsprechende SPSSX Programm ist:

```
TITLE »Beschreibung der 3-Clusterlösung des Complete-Linkage«
FILE HANDLE AFAMD / NAME - »AFAM.DAT«
GET FILE AFAMD
CLUSTER AKERNF AVERW AINW AGESIN
  /MEASURE - BLOCK
  /METHOD•• COMPLETE (FAMTYP)
  /PLOT^NONE
  /PRINT-NONE
  /SAVE*CLUSTER (4)
BREAKDOWN TABLES = AKERNF AVERW AINW AGESIN BY FAMTYP3
```

Durch die Anweisung COMPLETE (FAMTYP) wird festgelegt, daß die Ergebnisse des Complete-Linkage der Variablen FAMTYP zugewiesen werden sollen. Diese **Zuweisung** wird nur durchgeführt, wenn der Aufruf der SPSS-X Prozedur CLUSTER eine SAVE - Anweisung enthält. Die in unserem Beispiel enthaltene SAVE - Anweisung bewirkt, daß die Zugehörigkeit der Klassifikationsobjekte für eine 4-Clusterlösung in der Variablen FAMTYP4 zwischengespeichert wird. An den in der METHOD-Anweisung definierten Namen FAMTYP wird also in SPSS-X intern die Anzahl der Cluster angehängt. Diese weitere Spezifikation ist notwendig, da in der SAVEAnweisung mehrere Clusterlösungen zwischengespeichert werden können. Die Anweisung

```
SAVE = CLUSTER(2,6)
```

würde z.B. dazu führen, daß die Zugehörigkeit der Klassifikationsobjekte bei einer 2 - , 3 - , 4 - , 5 - und 6-Clusterlösung in den Variablen FAMTYP2, FAMTYP3, FAMTYP4, FAMTYP5 und FAMTYP6 zwischengespeichert wird.

Durch die Anweisung BREAKDOWN wird die SPSS-X Prozedur Breakdown (SPSS Inc. 1986: 372 - 384) aufgerufen. Die Variablen, die in die Analyse einbezogen werden sollen, werden in der TABLES - Anweisung definiert. Die nominalen Variablen, für die Mittelwerte und Standardabweichungen berechnet werden sollen, stehen hinter der BY-Anweisung, die quantitativen Variablen vor der BY-Anweisung. In unserem Beispiel bewirkt also die TABLES-Anweisung, daß für die Variable FAMTYP4 die Mittelwerte und Standardabweichungen in den Klassifikationsmerkmalen AKERNF, AVERW, AINW und AGESIN berechnet werden. Diese sind in der Tabelle 2.5-2 wiedergegeben.

In die Tabelle wurden wiederum die englischen Bezeichnungen, wie sie SPSS-X ausgibt, übernommen. Aus der Tabelle ist ersichtlich, daß sich die 4 Cluster vor allem in den Klassifikationsmerkmalen AKERNF und AVERW unterscheiden. Das erste Cluster, das am häufigsten auftritt, besteht durchschnittlich aus 3 Mitgliedern der Kernfamilie und einem Verwandten, das zweite aus ungefähr 8 Personen, von denen 7 der Kernfamilie und 1 Person der Verwandtschaft angehören. Das dritte Cluster setzt sich schließlich aus 8 Personen zusammen, von denen 3 der Kernfamilie und 5 der Verwandtschaft angehören. Das vierte Cluster bildet ein Rest-**Cluster.**

Bei der 3-Clusterlösung werden die Cluster 3 und 4 zusammengefaßt. Dadurch verschwindet das Restcluster.

Neben den Mittelwerten sollten auch die Standardabweichungen, die die Heterogenität innerhalb eines Clusters messen, betrachtet werden. So z.B. beträgt die Standardabweichung für das erste Cluster in dem Klassifikationsmerkmal AKERNF 1.5, d.h., daß sich in diesem Cluster mit einer sehr hohen Wahrscheinlichkeit Klassifikationsobjekte mit einer Anzahl von 1.5 und 4.5 Mitgliedern der Kernfamilie befinden.

Eine **detailierte Beschreibung** der Cluster erhält man, wenn anstelle der Mittelwerte und Standardabweichungen Kreuztabellen zwischen der Clusterzugehörigkeit und den Klassifikationsmerkmalen berechnet werden. Das entsprechende SPSS-X Programm in unserem Beispiel wäre:

```
CROSSTABS TABLES = AKERNF, AVERW, AINW, AGESIN BY FAMTYP3
OPTIONS 3,4
STATISTICS ALL
```

Zur Beschreibung der SPSS-X Prozedur siehe SPSS Inc. (1986: 336-352) und Kapitel 3. Die Tabelle 2.5-3 enthält die Ergebnisse der Kreuztabellierung der 3-Clusterlösung mit den Klassifikationsmerkmalen AKERNF und AVERW.

Aus der Tabelle ist ersichtlich, daß das Cluster 2 von den beiden anderen Clustern durch das Klassifikationsmerkmal AKERNF getrennt wird und **die** Cluster 1 und 2 von dem Cluster 3 durch das Klassifikationsmerk-

Tabelle 2.5-2:

Mittelwerte und Standardabweichungen in den Klassifikationsmerkmalen fiir die 3-Clusterldsung des Complete-Linkage

	Klassifikationsmerkmale:			
CLUSTER:	AKERNF		AVERW	
FAMTY	MEAN	STD.DEV	MEAN	STD.DEV
1 (n = 94)	3.19	1.48	.85	1.15
2 (n« 32)	7.43	1.58	1.09	1.08
3 (n = 31)	3.48	1.78	5.33	1.89
4 (n = 2)	1.00	0.00	8.00	0.00
CLUSTER:	AINW		AGESIN	
FAMTY	MEAN	STD.DEV	MEAN	STD.DEV
1 (n = 94)	.14	.37	.37	.87
2 fa-32)	.09	.29	.78	1.12
3 (n = 31)	.03	.18	.06	.25
4 (n = 2)	.50	.71	2.50	2.12

mal AVERW. Sowohl in dem Klassifikationsmerkmal AKERNF und AVERW besitzt das Cluster 3 eine zweigipfelige Verteilung, die zum Teil auf die vorausgehende Verschmelzung (von Cluster 3 und 4) zurückzufuhren ist.

Zusammenfassend empfiehlt sich für eine oder mehrere Clusterlösungen, sofern diese in Betracht kommen, folgende Informationen zu berechnen:

1. **Einzelfallstudie.** Es werden für einige Klassifikationsobjekte deren Clusterzugehörigkeit und ihre Ausprägungen in den Klassifikationsmerkmalen ausgelistet und untersucht. Die Auswahl kann beispielsweise zufällig erfolgen.
2. **Berechnung von Mittelwerten und Standardabweichungen** der Cluster in den Klassifikationsmerkmalen.
3. **Berechnung von Kreuztabellierungen** der Cluster mit den Klassifikationsmerkmalen.

Nach einer Analyse dieser Informationen müssen Entscheidungen für die weitere Analyse getroffen werden (vgl. Abbildung 1.2-1). Als Aus-

Tabelle 2.5-3:

Zusammenhang der 3-Clusterlösung und den KLassifikationsmerkmalen AKERNF und AVERW (Zeilenprozente)

AKERNF

Cluster:	1	2	3	4	5	6	7	8	9	10	12
1	17	19	17	26	15	5	0	0	0	0	0
2	0	0	0	0	9	19	25	31	3	9	3
3	15	30	12	9	21	6	6	0	0	0	0

AVERW

Cluster:	0	1	2	3	4	5	6	7	8	9
1	57	16	12	14	1	0	0	0	0	0
2	37	31	16	16	0	0	0	0	0	0
3	0	0	9	3	16	30	12	9	15	6

gangspunkt dieser Entscheidung kann beispielsweise die Beantwortung der Frage der inhaltlichen Interpretierbarkeit der Cluster, ob also die erzielte oder die erzielten Clusterlösungen Sinn machen, gemacht werden. Abhängig von der Beantwortung wird man sich vorläufig für eine bestimmte Clusterlösung entscheiden oder aber nach einer neuen Klassifikation suchen, indem inhaltlich begründetete neue Entscheidungen über die bisherigen Schritte getroffen werden.

In unserem Beispiel könnten z.B. die Cluster der 3-Clusterlösung folgendermaßen interpretiert werden:

Cluster 1: Kernfamilie im heutigen Sinn, die aus einem Ehepaar und einem Kind besteht

Cluster 2: Große Kernfamilie, die sich aus einem Ehepaar und mehreren Kindern zusammensetzt.

Cluster 3: Großfamilie mit einer beträchtlichen Anzahl von Verwandten.

Gegen diese Interpretation läßt sich einwenden, daß durch die Zusammenfassung der Variablen »Stellung im Haushalt« (vgl. Abschnitt 2.1) die eigentliche familiale Struktur verdeckt wird. So können sich z.B. das 2. und 3. Cluster nur durch eine Generationsverschiebung des Haushalts**Vorstandes** unterscheiden, da zu den Verwandten auch Enkel - und Schwiegerkinder, aber auch die Großeltern gezählt wurden. Bei dem 2. Cluster kann nun der Hof bereits an den Sohn übergeben worden sein, der folglich mit seiner Gattin (Schwiegertochter der Eltern des Sohnes) und seinen Kindern die Kernfamilie bildet. Beim 3. Cluster dagegen hat noch keine Hofübergabe stattgefunden. Die Schwiegertochter bzw. der Schwiegersohn und die Enkelkinder des Haushaltsvorstandsehepaares bilden deshalb die Verwandten. Wird dieser Einwand akzeptiert, wird man eine neue Clusteranalyse mit einer feineren Differenzierung der familialen Haushaltsstruktur durchführen, indem die ursprünglich 26 Ausprägungen der Variablen »Stellung im Haushalt« in eine größere Anzahl von Dummy-Variablen aufgelöst wird.

Ein anderer **Ausgangspunkt** für die Entscheidung über das weitere Vorgehen wäre, die Frage der **Stabilität** der erzielten Clusterlösung vor die inhaltliche Gültigkeitsprüfung zu stellen. (Die Schritte 8 und 9 der Abbildung 1.2-1 werden also vertauscht). Die Stabilität einer Clusterlösung stellt ein formales Kriterium dar und meint, daß bei ähnlichen Daten und/oder ähnlichen Datentransformationen und/oder ähnlichen Unähnlichkeitsmessungen und/oder ähnlichen Clusteranalyseverfahren ähnliche oder idealerweise identische Ergebnisse erzielt wird, wobei »ähnlich« weiter zu präzisieren ist. Aber auch wenn die Entscheidung zunächst von der inhaltlichen Fragestellung der Interpretierbarkeit abhängig gemacht wird, muß bei einer inhaltlich interpretierbaren Lösung die Stabilität überprüft werden, um Fehlinterpretationen, die auf Artefakten beruhen, zu vermeiden. Verfahren der Stabilitätsprüfung werden im nächsten Kapitel behandelt.

2.6 Besonderheiten der SPSS-X Prozedur CLUSTER

Die SPSS-X Prozedur CLUSTER weist mehrere Besonderheiten auf, die an dieser Stelle zusammengefaßt werden sollen:

1. Die Klassifikationsobjekte werden **fortlaufend** mit 1 beginnend **nummeriert.** Eine **Identifikation** ist deshalb nur durch die Verwendung einer **String-Variablen** möglich, bei der den Klassifikationsobjekten Namen gegeben werden.

2. Das Verschmelzungsniveau wird bei dem **Dendrogramm** auf das Intervall 0 bis 25 **reskaliert,** wobei auch Unähnlichkeiten bzw. Ähnlichkeiten mit einem Wert von 0 einen Wert größer 0 erhalten, damit die Verschmelzung noch gezeichnet werden kann.

3. **Ähnlichkeitsmaße** werden **nicht** in Unähnlichkeitsmaße **transformiert,** sondern der Algorithmus wird auf Ähnlichkeitsmaße angewendet.

4. In der METHOD - Anweisung können **mehrere Verfahren** definiert werden. In älteren SPSS-X Versionen hat diese wiederholte Anwendung zu Fehlern geführt. Es ist deshalb sinnvoll die zur Verfügung stehende SPSS-X Version auf diesen Fehler hin zu überprüfen.

5. Die hierarchisch agglomerativen Verfahren können auch für eine **Analyse von Klassifikationsmerkmaien** eingesetzt werden. Technisch bedeutet das nur, daß die Klassifikationsdatenmatrix umgedreht (transponiert) wird. Diese Operation ist in SPSS-X nicht als Prozedur vorsehen. Soll aber z.B. die Korrelation zwischen den Klassifikationsmerkmalen als Ähnlichkeitsmaß verwendet werden, dann ist das Vorgehen denkbar einfach. In einem ersten Schritt der Analyse wird eine Korrelationsmatrix der Klassifikationsmerkmale mit der Prozedur PEARSON CORR (SPSS Inc. 1986: 638-646) berechnet und zwischengespeichert. Sie wird dann in der anschließenden CLUSTER-Prozedur als Ähnlichkeitsmatrix eingelesen und analysiert. In unserem Beispiel würde sich folgendes SPSS-X Programm ergeben:

```
FILE HANDLE AFAMD/NAME - »AFAM.DAT«
FILE HANDLE AKORR/NAME - »AKORR.DAT«
GET FILE - AFAMD
PROCEDURE OUTPUT «= AKORR
PEARSON CORR AKERNF AVERW AINW AGESIN
OPTIONS 4
INPUT MATRIX FILE - AKORR
CLUSTER AKERNF AVWER AINW AGESIN
     /READ - SIMILAR
     /..
```

Durch die Anweisungen FILE HANDLE AKORR/NAME - »AKORR.DAT« und PROCEDURE OUTPUT - AKORR wird festgelegt, daß die SPSS-X interne Datei AKORR auf der externen Datei AKORR.DAT zwischengespeichert wird. Auf diese Datei wird durch die OPTIONS - Anweisung bei der Prozedur PEARSON CORR die Korrelationsmatrix der Klassifikationsmerkmale gespeichert. Die Korrelationsmatrix wird nun unmittelbar durch die Anweisung INPUT MATRIX FILE = AKORR eingelesen. In der SPSS-X Prozedur CLUSTER muß nun definiert werden, daß eine Ähnlichkeitsmatrix einge-

lesen wird. Das geschieht durch die Anweisung READ = SIMILAR. Sollen andere Un- oder Ähnlichkeitsmaße verwendet werden, kann anstelle der Prozedur PEARSON CORR die Prozedur PROXIMITIES (SPSS Inc. 1986: 732-750) verwendet werden.

6. In der SPSS-X-Prozedur ist nur die **fallweise Elimination von fehlenden Werten** möglich (s. dazu Kapitel 4.1).

3. Verfahren der Stabilitätsprüfung

3.1 Die Logik der Stabilitätsprüfling

Die Vorstellung der **Stabilität** einer Clusterlösung geht über das Konzept der **Zuverlässigkeit** der klassischen Testtheorie **hinaus. Zuverlässigkeit** einer Messung, z.B. eines Intelligenztests in der Psychologie oder einer Einstellungsskala in der Sozialpsychologie oder der Soziologie, liegt dann vor, wenn ein Meßinstrument (der Intelligenztest oder die Einstellungsskala) unter **identischen Meßbedingungen** im Idealfall identische aber doch zumindest sehr ähnliche Ergebnisse liefert. Die Meßbedingungen dürfen sich dabei nur durch **zufällige Störgrößen** unterscheiden. Die für die Zuverlässigkeitsprüfung entwickelten statistischen Verfahren, wie z.B. die Faktorenanalyse oder die Split-Half-Analyse, ermöglichen nur eine Schätzung dieser zufälligen Meßfehler.

Systematische Meßfehler, die in der quantitativen Geschichtsforschung beispielsweise durch das Abschreiben von Quellen oder durch bewußte Unter - oder Überschätzungen in staatlichen Statistiken entstehen, können nicht erfaßt werden.

In die Klassifikationsanalyse wurde das Zuverlässigkeitskonzept und einige Prüfverfahren übernommen oder neuentwickelt. Als Beispiele seien hier nur erwähnt:

- **Die Sensitivitätsanalyse** (Blashfield u.a. 1982:173): Bei der Sensitivitätsanalyse werden die empirisch beobachteten Ausprägungen der Klassifikationsobjekte in den Klassifikationsmerkmalen mit einem zufälligen Meßfehler überlagert. Daran anschließend wird eine Clusteranalyse für die ursprünglichen und für die mit einem Zufallsfehler überlagerten Ausprägungen berechnet und die beiden Ergebnisse miteinander verglichen. Das Verfahren der Sensitivitätsanalyse wird in Abschnitt 3.3.3 ausführlich dargestellt.
- **Split-Half-Verfahren:** Die Klassifikationsobjektc werden zufällig in zwei Stichproben zerlegt. Daran anschließend wird für jede Unter-Stichprobe eine getrennte Clusteranalyse durchgeführt und ein Vergleich der beiden Ergebnisse durchgeführt.

Die Zuverlässigkeitsprüfung setzt identische Meßbedingungen, die nur durch Zufallsfehler überlagert sind, voraus. Bei der **Stabilitätsprüfung** dagegen werden die **Meßbedingungen systematisch variiert,** um Anhaltspunkte über die »Allgemeingültigkeit« (Stabilität) der erzielten Clusterlösung zu erhalten. Die Meßbedingungen umfassen dabei alle Entschei-

düngen (Schritte) zur Lösung eines Klassifikationsproblems. Diese sind (vgl. Kapitel 1):

- die Auswahl der Klassifikationsobjekte,
- die Auswahl der Klassifikationsmerkmale,
- die Behandlung fehlender Werte, sofern solche vorliegen,
- die Transformation der Klassifikationsmerkmale, sofern eine solche durchgeführt wurde,
- die Auswahl des Ähnlichkeits - oder Unähnlichkeitsmaßes und
- die Auswahl des Clusteranalyseverfahrens.

Insbesondere wird man jene Bedingungen variieren, für die keine eindeutigen inhaltlich begründeten Entscheidungen getroffen werden konnten, z.B. bezüglich des verwendeten Verfahrens.

Sowohl bei den Verfahren der Zuverlässigkeits- als auch bei denen der Stabilitätsprüfung muß die Ähnlichkeit oder Unähnlichkeit der erzielten Clusterlösungen für die unterschiedlichen Meßbedingungen berechnet werden. Grundsätzlich können die erzielten Clusterlösungen hinsichtlich

- der durch die **Cluster»nalyse berechneten Ähnlichkeits-** oder **Unähnlichkeitsmatrix** der Klassifikationsobjekte (Berechnung von Matrixkorrelationen),
- der **Zuordnung der Klassifikationsobjekte zu den Clustern** (Berechnung von Kreuztabulierungen) und/oder
- **der Verteilung der Klassifikationsmerkmale** in den berechneten Clustern oder von Charakteristiken dieser Verteilungen, wie z.B. den Clusterm ittel werten

miteinander verglichen werden. Diese drei Verfahren werden im folgenden kurz dargestellt.

3.2 Die Berechnung der Unähnlichkeit bzw. Ähnlichkeit von Clusterlösungen

3.2.1 Berechnung von Matrixkorrelationen

Bei der Berechnung von Matrixkorrelationen werden die für unterschiedliche Meßbedingungen durch die Clusteranalyse berechneten Unähnlichkeits - oder Ähnlichkeitsmatrizen der Klassifikationsobjekte miteinander verglichen. Die Berechnung von Matrixkorrelationen läßt sich in SPSS-X nur schwer realisieren und soll deshalb nur anhand des fiktiven Beispiels aus dem Abschnitt 2.4.1 dargestellt werden. Ausgangspunkt der Berechnung bildet das Verschmelzungsschema. Aus diesem läßt sich abhängig davon, ob Unähnlichkeiten oder Ähnlichkeiten zwischen den Klas-

sifikationsobjekten in die Analyse einbezogen wurden, eine **»theoretische« Ähnlichkeits** - oder **Unähnlichkeitsmatrix** berechnen.

Tabelle 3.2-1:

Verschmelzungsschema des Complete-Linkage für das fiktive Rechenbeispiel des Abschnitts 2.4

Schritt	Cluster 1	Cluster 2	Verschmelzungsniveau
1	A	B	2.0
2	A,B	C	3.0
3	D	E	5.0
4	A,B,C	D,E	10.0

Betrachten wir beispielsweise das Verschmelzungsschema des Complete-Linkage für das fiktive Rechenbeipiel des Abschnittes 2.4.1 (vgl. Tabelle 3.2-1).

Der Complete-Linkage wurde für eine Unähnlichkeitsmatrix berechnet. In dem Verschmelzungsschema beträgt die »theoretische« Unähnlichkeit zwischen A und B 2.0, die zwischen A und C 3.0 und die zwischen B und C ebenfalls 3.0, da A und B ein Cluster bilden. Die »theoretische« Unähnlichkeit zwischen D und E beträgt 5.0, die zwischen A und D 10.0, zwischen A und E 10.0, usw. Diese Werte können unmittelbar in die theoretische Unähnlichkeitsmatrix eingetragen werden (vgl. Tabelle 3.3-2).

Tabelle 3.2-2:

»Theoretische« Unähnlichkeitsmatrix des Complete-Linkage für das fiktive Rechenbeispiel des Abschnittes 2.4

Kl.objekt	A	B	C	D	E
A	0.0				
B	2.0	0.0			
C	3.0	3.0	0.0		
D	10.0	10.0	10.0	0.0	
E	10.0	10.0	10.0	5.0	0.0

Zur Berechnung der theoretischen Unähnlichkeiten (oder Ähnlichkeiten) braucht also das Verschmelzungsschema nur von oben nach unten gelesen und dabei die entsprechenden Unähnlichkeiten (oder Ähnlichkeiten) notiert werden.

Tabelle 3.2-3:

»Theoretische« Unähnlichkeitsmatrix des Single-Linkage für das fiktive Rechenbeispiel des Abschnittes 2.4

Kl.objekt	A	B	C	D	E
A	0.0				
B	2.0	0.0			
C	2.5	2.5	0.0		
D	6.5	6.5	6.5	0.0	
E	6.5	6.5	6.5	5.0	0.0

Nach demselben Vorgehen kann die »theoretische« Unähnlichkeitsmatrix für den Single-Linkage für das fiktive Rechenbeispiel gebildet werden. Diese ist in der Tabelle 2.4-3 dargestellt.

Zur Berechnung der Ähnlichkeit von Ähnlichkeits oder Unähnlichkeitsmatrizen haben Rolf und Sokal (Sneath & Sokal 1972: 277-280) die Anwendung der einfachen Produkt-Moment-Korrelation (Pearsons Korrelation) vorgeschlagen. In die Berechnung wird das untere Dreieck (ohne Diagonale) der Unähnlichkeitsmatrizen einbezogen. Jede Zelle dieses unteren Dreieckes stellt dabei eine Beobachtung dar. Der Rechenvorgang ist in der Tabelle 3.2-4 abgebildet.

In dem fiktiven Beispiel beträgt die Matrixkorrelation zwischen Single- und Complete-Linkage 0.98. Die beiden Clusterergebnisse stimmen weitgehend überein. Ein eindeutiges Kriterium, ab dem die Matrixkorrelation als »befriedigend« gilt, fehlt allerdings.

Eine Matrixkorrelation kann auch zwischen der »theoretischen« und der »empirischen« Unähnlichkeitsmatrix oder Ähnlichkeitsmatrix berechnet werden. Die Matrixkorrelation wird in diesem Fall als kophenetische Korrelation bezeichnet.

Entscheidend für den Unterschied zu den anderen beiden in 3.1 dargestellten Verfahren ist die Tatsache, daß bei der Matrixkorrelation die gesamte Information des Verschmelzungsprozesses eingeht. In die beiden anderen Verfahren dagegen geht eine bestimmte Clusterlösung. z.B. eine 3-Cluster- oder eine 4-Clusterlösung, ein.

Übungsaufgabe 9: Berechnen Sie für die Aufgabe 4a die Matrixkorrelation zwischen Complete- und Single Linkage.

Tabelle 3.2-4:

Rechenschema zur Berechnung der Matrixkorrelation für das fiktive Beispiel

Zelle	Korrelation zwischen dem Complete	Single			
i	C_i	S_i	C_i^2	S_i^2	C_iS_i
1	2.0	2.0	4.0	4.0	4.0
2	3.0	2.5	9.0	6.25	7.5
3	3.0	2.5	9.0	6.25	7.5
4	10.0	6.5	100.0	42.25	65.0
5	10.0	6.5	100.0	42.25	65.0
6	10.0	6.5	100.0	42.25	65.0
7	10.0	6.5	100.0	42.25	65.0
8	10.0	6.5	100.0	42.25	65.0
9	10.0	6.5	100.0	42.25	65.0
10	5.0	5.0	25.0	25.00	25.0
Σ	73.0	51.0	647.0	295.00	434.0

$S_{CC} = 647 - 73.0*73.0/10 = 114.1$
$S_{SS} = 295 - 51.0*51.0/10 = 34.9$
$S_{CS} = 434 - 73.0*51.0/10 = 61.7$
$CORR(C,S) = 61.7/(114.1*34.9)^{1/2} = 0.978$

3.2.2 Ähnlichkeit auf der Grundlage der Zuordnung der Klassifikationsobjekte zu den Clustern

Dieses Verfahren läßt sich in SPSS-X am einfachsten von den drei Verfahren realisieren. Das allgemeine Vorgehen besteht darin, daß für unterschiedliche Ausgangskonstellationen (Meßbedingungen) die Clusterzugehörigkeit der Klassifikationsobjekte berechnet und zwischengespeichert wird. In einem weiteren Schritt der Analyse werden die Ergebnisse für diese unterschiedlichen Ausgangskonstellationen durch einfache Kreuztabulierungen miteinander verglichen.

In unserem empirischen Beispiel der familialen Haushaltsstrukturen ergibt sich für den Complete- und Single-Linkage für eine 3-Clusterlösung die in der Tabelle 3.2-5 dargestellte Kreuztabelle.

Aus der Tabelle läßt sich erkennen, daß beim Single-Linkage nur ein großes Cluster, das 157 der 159 Haushalte enthält, gebildet wird. Deshalb wird die Übereinstimmung (Ähnlichkeit) der beiden Clusterlösungen sehr gering sein.

Tabelle 3.2-5:

Zusammenhang zwischen der 3-Clusterlösung des Complete- und Single-Linkage

Complete - linkage:	Single-Linkage: Cluster 1	Cluster 2	Cluster 3	Gesamt
Cluster 1	93	1	0	94
Cluster 2	32	0	0	32
Cluster 3	32	0	1	33
Gesamt	157	1	1	159

Zur Messung der Ähnlichkeit der beiden Clusterlösungen können alle symmetrischen Assoziationskoeffizienten für nominale Variablen verwendet werden. In der SPSSX-Prozedur CROSSTABS (SPSS Inc. 1986: 336-352), mit der Tabellenanalysen durchgeführt werden können, stehen folgende Zusammenhangsmaße zur Auswahl:

- das symmetrische Lambda,
- der symmetrische Unsicherheitskoeffzient,
- der Kontingenzkoeffizient und
- Cramers V.

Insbesondere empfiehlt sich die Verwendung des Lambda Koeffizienten, da diesem ein sehr einfaches Konzept zugrundeliegt. Der Koeffizient Lambda h wurde ursprünglich zur Messung kausaler, also asymmetrischer Beziehungen zwischen zwei nominalen Variablen entwickelt. Er gehört der Klasse der PRE-Koeffizienten (Proportional Reduction of Errors) an. Diese geben an, um wieviele Prozentpunkte sich die Prognose der abhängigen Variablen verbessert, wenn die unabhängigen Variablen in die Prognose der abhängigen Variablen einbezogen werden, im Vergleich zu einer Prognose der abhängigen Variablen ohne Berücksichtigung der unabhängigen Variablen.

Bezeichnen wir mit X_1, X_2,.. die unabhängigen Variablen und mit Y die abhängige Variable, dann ist der PRE-Koeffizient allgemein definiert als:

$$PRE = \frac{\text{Prognosefehler ohne } X_1,X_2,.. \; - \; \text{Prognosefehler mit } X_1,X_2,..}{\text{Prognosefehler ohne } X_1,X_2,...}$$

Da bei Lambda h nur eine unabhängige Variable vorliegt, vereinfacht sich der PRE-Koeffizient zu

$$PRE = h_{YX} = \frac{\text{Prognosefehler ohne X} \; - \; \text{Prognosefehler mit X}}{\text{Prognosefehler ohne X}}$$

und falls X die abhängige Variable ist, zu

$$PRE = h_{XY} = \frac{\text{Prognosefehler ohne Y} \; - \; \text{Prognosefehler mit Y}}{\text{Prognosefehler ohne Y}}$$

Für unser Beispiel sind diese Prognosefehler, wobei X die Ergebnisse des Complete-Linkage und Y die des Single-Linkage sein sollen:

Prognosefehler ohne X: Die beste Prognose für Y ohne Kenntnis von X ist die Voraussage, daß ein Klassifikationsobjekt dem ersten Cluster angehört. Durch diese Prognose werden 157 richtige und 2 falsche Prognosen gemacht. Der Prognosefehler beträgt also 2. Die Voraussage eines anderen Clusters würde zu wesentlich höheren Prognosefehlem führen. (Bei der Definition des PRE Koeffizienten müßte eigentlich an Stelle von »Prognosefehler« »minimalster« Prognosefehler geschrieben werden.)

Prognosefehler mit X: Für jede Ausprägung von X wird getrennt eine Prognose von Y errechnet. Für die erste Ausprägung von X (erstes Cluster des Complete-Linkage) ist die beste Voraussage das Cluster 1 des Single-Linkage (93 richtige Prognosen und 1 falsche Prognose), für die zweite Ausprägung von X (zweites Cluster des Complete Linkage) ebenfalls das Cluster 1 des Single-Linkage (32 richtige Prognosen und 0 falsche Prognosen) und für die dritte Ausprägung von X (drittes Cluster des Complete-Linkage) ebenfalls das Cluster 1 des Single-Linkage (32 richtige Prognosen und 1 falsche Prognose). Das Aufsummieren dieser einzelnen Prognosefehler ergibt insgesamt einen Prognosefehler von 2.

Damit ist Lambda h_{YX} 0.0. Die Ergebnisse des Complete-Linkage für drei Cluster leisten keinen Beitrag zur Prognose der Ergebnisse des Single-Linkage. Im Idealfall müßte h_{YX} 1.0 sein.

Analog kann Lambda h_{xy} berechnet werden. Es beträgt für unser Beispiel 0.015. Die Ergebnisse des Single-Linkage tragen also auch kaum zur Prognose (Erklärung) der Ergebnisse des Complete-Linkage bei. Die von uns für eine City-Blockmetrik berechnete 3-Clusterlösung besitzt also keine Stabilität hinsichtlich des Single- und Complete-Linkage.

Aus den berechneten Prognosefehlern kann als Gesamtmaß das symmetrische Lambda $sym.h_{xy}$ berechnet werden:

$$sym.h_{XY} = \frac{\text{Prognosefehler ohne Y} - \text{Prognosefehler mit Y} + \text{Prognosefehler ohne X} - \text{Prognosefehler mit X}}{\text{Prognosefehler ohne Y} + \text{Prognosefehler ohne X}}$$

In unserem Beispiel ist $sym.h_{xy} = 0.0149$, was nur die obige Interpretation bestätigt.

Single- und Complete-Linkage besitzen aber nun extreme, einander entgegengesetzte Eigenschaften. Deshalb kann es sinnvoll sein, ein weiteres Verfahren zu verwenden, das in der Mitte dieser beiden Verfahren liegt, also z.B. das Baverage-Verfahren in SPSSX. Daneben kann noch eine Variation des Metrikparameters sinnvoll sein. Die Tabelle 3.2-6 enthält die Ergebnisse, die für diese Ausgangskonstellationen erzielt werden.

Tabelle 3.2-6:

Ergebnisse der Stabilitätsprüfung der familialen Haushaltsdaten (symmetrische Lambdas)

Metrik	Verfahren	Block			Euclid		
		C	S	B	C	S	B
Block	Complete	**0**					
	Single	**.01**	**0**				
	Bavcragc	**.27**	**.03**	**0**			
Euclid	Complete	**.51**	**.02**	**.22**	**0**		
	Single	**.01**	**1.00**	**.03**	**.02**	**0**	
	Baverage	**.17**	**.06**	**.25**	**.40**	**.07**	**0**

Nur für den Single-Linkage wird für beide Unähnlichkeitsmaße eine perfekte Übereinstimmung erzielt. Inhaltlich könnte das Ergebnis des Single-Linkage, bei dem ein sehr großes Cluster gebildet wird, bedeuten, daß der Großteil der Haushalte Realisierungen desselben familialen Prozesses darstellen und sich die Haushalte nur in unterschiedlichen Phasen befin-

den. Für den Complete-Linkage beträgt die Übereinstimmung für beide Unähnlichkeitsmaße 0.51, für den BaverageLinkage sogar nur 0.40. Insgesamt stimmen die Ergebnisse des Complete- und BaverageLinkage besser überein als die des Single-Linkage und den beiden anderen Verfahren.

Abschließend soll noch das SPSS-X Programm dargestellt werden, das die in der Tabelle 3.2-6 enthaltenen symmetrischen Lambda-Koeffizienten berechnet.

```
TITLE »STABILITAETSPRUEFUNG DER CLÜSTERLOESUNG «
FILE HANDLE AFAMD / NAME - »AFAM.DAT«
GET FILE AFAMD
CLUSTER AKERNF AVERW AINW AGESIN
  /MEASURE - BLOCK
  /METHOD - SINGLE(S1) COMPLETE(Cl) BAVERAGE(Bl)
  /PRINT « NONE
  /PLOT - NONE
  /SAVE - CLUSTER(3)
CLUSTER AKERNF AVERW AINW AGESIN
  /MEASURE - EUCLID
  /METHOD ~ SINGLE(S2) COMPLETE(C2) BAVERAGE(B2)
  /PRINT - NONE
  /PLOT - NONE
  /SAVE - CLUSTERS)
CROSSTABS TABLES - CI3 BY SI 3 Bl 3 C23 S23 B23/
                   SI3 BY B13 C23 S23 B23/
                   B13 BY C23S23 B23/
                   C23 BY S23 B23/
                   B23 BY S23
STATISTICS 4
```

Beim Aufruf der ersten CLUSTER - Prozedur werden die 3-Clusterlösungen für den Single - , Complete und BaverageLinkage für die City-Blockmetrik berechnet und in den Variablen S13, C13 und B13 zwischengespeichert. Die Anweisungen PLOT= NONE und PRINT= NONE bewirken, daß eine Ausgabe des Eiszapfendiagramms (Voreinstellung der PLOTAn**Weisung)** und des Verschmelzungsschemas (Voreinstellung der PRINTAnweisung) unterdrückt wird.

Durch den zweiten Aufruf der CLUSTER - Prozedur werden analog die 3-Clusterlösungen für die euklidische Metrik zwischengespeichert.

Durch die Anweisung CROSSTABS wird die SPSS X Prozedur CROSSTABS aufgerufen. Die zu berechnenden Tabellen werden in der TABLES - Anweisung vereinbart. Die Anweisung TABLES = C13BYS13 B13 C23 S23 B23 bewirkt also z.B., daß die 2 - dimensionalen Tabellen zwischen der 3-Clusterlösung des Complete-Linkage für die City-Blockmetrik (C13) mit der 3-Clusterlösung des Single-Linkage für die City-Blockmetrik (S13), zwischen C13 und der 3-Clusterlösung des Baverage-Linkage für die City-Blockmetrik (B13) usw. berechnet werden.

Durch die Anweisung STATISTICS 4 werden als Zusammenhangsmaße nur die Lambda-Koeffizienten berechnet.

Hinwels: In filteren SPSS-X-Versionen hat die Anwendung mehrerer Clusterverfahren in einer METHOD Anweisung zu Berechnungsfehlern geführt. Es empfiehlt sich deshalb, zunächst durch getrennte Clusteranalysen die zur Verfügung stehende SPSS X Version auf diesen Fehler hin zu überprüfen.

Übungsaufgabe **10:** Für den Single - und Baverage Linkage ergibt sich in unserem Beispiel folgende Kreuztabelle:

	Baverage-Linkage Cluster 1	Cluster 2	Cluster 3
Single-Linkage			
Cluster 1	123	33	1
Cluster 2	1	0	0
Cluster 3	0	1	0

Berechnen Sie die asymmetrischen und symmetrischen Lambda Koeffizienten und interpretieren Sie diese!

3.2.3 Ähnlichkeit aufgrund der Verteilung der Klassifikationsmerkmale in den Clustern

Formal bilden die berechneten Cluster Aggregate, die i.d.R. durch eine Verteilung in den Klassifikationsmerkmalen gekennzeichnet sind. In Abschnitt 2.4.5 wurde bereits dargestellt, wie Klassifikationsobjekte dieser Gruppe behandelt werden können. Die Durchführung soll nun in SPSS X anhand unserer Daten beschrieben werden.

Berechnug der Unähnlichkeit zwischen Clusterlösungen auf der Grundlage der Clustermittelwerte in den Klassifikationsmerkmalen:

Das Vorgehen besteht aus zwei Schritten:

Schritt **1:** Berechnen und Abspeichern der Clustermittelwerte für unterschiedliche Ausgangsbedingungen (Clusterlösungen).

Schritt **2:** Berechnen der Unähnlichkeit der Clusterlösungen.

In unserem Beispiel wird der erste Schritt durch folgendes SPSS X Programm gelöst:

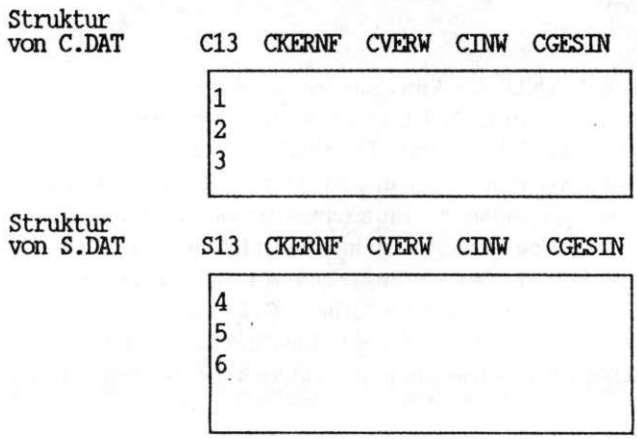
```
FILE HANDLE AFAMD / NAME = » AFA M.DAT«
GET FILE AFAMD
CLUSTER AKERNF AVERW AINW AGESIN
  /METHOD - COMPLETE (CD SINGLE (Sl)
```

```
    /MEASURE - BLOCK
    /PLOT - NONE
    /PRINT - NONE
    /SAVE - CLUSTERO)
  FILE HANDLE CDAT/NAME - »CDAT«
  AGGREGATE OUTFILE - CDAT
    /BREAK - C13
    /CKERNF CVERW CINW CGESIN -
    MEAN(AKERNF,AVERW,AINW,AGESIN)
  RECODE S13 (1-4) (2-5) (3-6)
  FILE HANDLE SD AT/N AM E - »S.DAT«
  AGGREGATE OUTFILE - SDAT
    /BREAK - S13
    /CKERNF CVERW CINW CGESIN -
    MEAN(AKERNF,AVERW,AINW,AGESIN)
```

In der SPSS-X Prozedur CLUSTER werden zunächst die 3-Clusterlösungen des Single - und Complete-Linkage für die City-Blockmetrik berechnet und in der SPSS-X Arbeitsdatei unter den Variablen C13 und SP zwischengespeichert. Durch die beiden folgenden AGGREGATE - Anweisungen werden die Clustermittel werte berechnet und extern abgespeichert. (Die Mittelwerte der 3-Clusterlösung des Complete-Linkage werden auf der externen SPSS-X Systemdatei CDAT unter den Variablen CKERNF, CVERW, CINW und CGESIN abgespeichert, die Mittelwerte des Single-Linkage auf der externen SPSS-X Systemdatei S.DAT.) Die Dateien CDAT und S.DAT müssen zuvor durch eine FILE HANDLE - Anweisung definiert werden.

Abbildung 3.2-1:

Struktur von CDAT und S.DAT



Die RECODE - Anweisung vor Aufruf des zweiten AGGREGATE bewirkt, daß die Ausprägungen der Clusterzugehörigkeit des Single Linkage umkodiert werden. Dadurch erhält man für den nächsten Schritt eine fortlaufende Identifikationsvariable.

Abbildung 3.2-1 zeigt die Struktur der beiden externen Dateien CDAT und S.DAT. Die restlichen Felder der Datenmatrizen enthalten die entsprechenden Mittelwerte.

Die Anordnung der AGGREGATE - Prozeduren ist beliebig. Zu beachten ist nur, daß identische Variablenbezeichnungcn verwendet werden. Dadurch können aufwendige RENAME Befehle im nachfolgenden Schritt vermieden werden. In unserem Beispiel muß die Variable S13 oder C13 umbenannt werden. Zweitens muß eine Rekodierung einer der beiden Variablen C13 oder B13 durchgeführt werden.

Im zweiten Schritt werden nun diese beiden Dateien eingelesen und die Unähnlichkeiten der beiden Clusterlösungen berechnet.

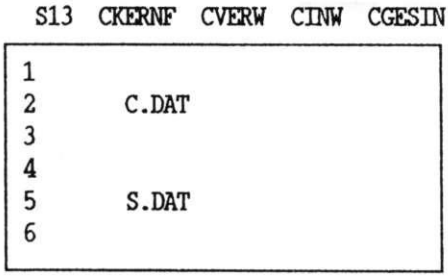
```
FILE HANDLE CDAT/NAME - »CDAT«
FILE HANDLE SDAT/NAME - »S.DAT«
ADD FILES FILE - CDAT/RENAME - (C13 - S13)
  /FILE - SDAT
LIST VARIABLES - S13, CKERNF, CVERW, CINW, CGESIN
STR1NG CL (A20)
RECODE S13 ( 1 - t n
  (2*'C2')
  (3 = 'C3')
  (4-'SD
  (5«'S2')
  (6-'S3) INTO CL
CLUSTER CKERNF CVERW CINW CGESIN
  /MEASURE - BLOCK
  /PRINT - DISTANCE
  /PLOT - NONE
  /ID - CL
```

Durch die FILE HANDLE - Anweisungen wird die Verbindung zwischen den externen Dateien CDAT und S.DAT und den SPSS X internen Dateien CDAT und SDAT hergestellt. Die ADD FILES - Anweisung bewirkt, daß die beiden Dateien CDAT und SDAT zu einer SPSS X Arbeitsdatei verbunden werden, indem sie hintereinander zusammengefügt werden. Der Aufbau der neuen SPSS XSystemdatei ist in der Abbildung 3.2-2 dargestellt. Die RENAME - Anweisung nach Aufruf der Datei CDAT durch FILE = CDAT bewirkt, daß die Variable C13 ebenfalls den Namen S13 erhält. Damit ist S13 eine fortlaufende Identifikationsvariable.

Die LIST VARIABLES - Anweisung bewirkt, daß die Variable C13 und die Clustermittelwerte in den Variablen CKERNF, CVERW, CINW und CGESIN ausgedruckt werden.

Abbildung 3.2-2:

Aufbau der aus den Dateien CDAT und S.DAT zusammengefügten SPSS-X Datei



Durch die Anweisung STRING CL (A20) wird eine Stringvariable definiert, deren Ausprägungen aus 20 Zeichen bestehen können. Sie kann in der anschließenden CLUSTER Prozedur zur Benennung der Fälle verwendet werden. Durch die RECODE - Anweisung erhält die erste Ausprägung der Variablen S13 (erstes Cluster des Complete-Linkage) die Ausprägung Cl in der Stringvariablen CL, die zweite Ausprägung von S13 (zweites Cluster des Complete Linkage) die Ausprägung C2, usw...

In der CLUSTER - Prozedur werden schließlich die Unähnlichkeiten zwischen den Clustern berechnet. Die Anweisungen PRINT = DISTANCE und PLOT = NONE führen dazu, daß nur die Distanzmatrix zwischen den Clustern ausgegeben wird. Zur Messung der Unähnlichkeit wird die City-Blockmetrik verwendet. Für unsere Daten erhält man folgende Ergebnisse (siehe Abbildung 3.3-3).

Wie in der Abbildung ersichtlich ist, besitzen alle Cluster des Complete Linkage zum 1. Cluster des Single-Linkage den größten absoluten Abstand und zum 2. Cluster des Single Linkage den kleinsten absoluten Abstand. Im Idealfall sollte das Ergebnis die in der Abbildung 3.2-3a dargestellte Struktur besitzen.

In dem fiktiven Ergebnis sind sich das erste Cluster des Complete-Linkage und das erste des Single-Linkage, das zweite Cluster des Complete-Linkage und das zweite des Single Linkage und das dritte des Complete-Linkage und des Single-Linkage vollkommen identisch. Während sich die Cluster untereinander sehr deutlich unterscheiden.

Abbildung 3.2-3:

Distanzmatrix zwischen den Clustermittelwerten der 3-Clusterlösungen des Single- und Complete-Linkage

	C1	C2	C3	S1	S2	S3
C1	0					
C2	4.9	0				
C3	5.0	9.1	0			
S1	13.1	16.6	8.7	0		
S2	1.9	4.4	4.7	13.0	0	
S3	6.4	8.4	6.9	12.0	4.8	0

Abbildung 3.2-3a:

Idealstruktur der Distanzmatrix bei Stabilität

	C1	C2	C3	S1	S2	S3
C1	0					
C2	10	0				
C3	20	30	0			
S1	0	10	20	0		
S2	10	0	20	12	0	
S3	18	12	0	30-	28	0

Berechnung der Unähnlichkeit auf der Grundlage der Verteilung der Cluster in den Klassifikationsmerkmalen:

Sollen in die Berechnung der Unähnlichkeit von Clusterlösungen die Verteilungen der Klassifikationsmerkmale einbezogen werden, müssen die Klassifikationsmerkmale in Dummy-Variablen aufgelöst werden. Da die Klassifikationsmerkmale quantitatives Meßniveau besitzen, ist die Verwendung von ordinalen Dummy-Variablen sinnvoll. Wir beschreiben hier

nur den ersten Schritt des Vorgehens. Der zweite Schritt verläuft analog wie oben dargestellt.

```
FILE HANDLE AFAMD/NAME - »AFAM.DAT«
GET FILE AFAMD
CLUSTER AKERNF AVERW AINW AGESIN
  /METHOD - COMPLETE (Cl) SINGLE (SI)
  /MEASURE - BLOCK
  /PLOT - NONE
  /PRINT - NONE
  /SAVE - CLUSTERS)
DO REPEAT K - Kl TO K20 / WERT - 1 TO 20
COMPUTE K = 0
IF (AKERNF GE WERT) K« 1
END REPEAT
DO REPEAT V - VI TO V20 / WERT = 1 TO 20
COMPUTE V-0
IF (AVERW GE WERT) V- 1
END REPEAT
DO REPEAT I - II TO 120 / WERT - 1 TO 20
COMPUTE 1-0
IF (AINW GE WERT) I~1
END REPEAT
DO REPEAT G = Gl TO G20 / WERT - 1 TO 20
COMPUTE G = 0
IF (AGESIN GE WERT) G = 1
END REPEAT
FILE HANDLE CDAT/NAME - »C.DAT«
AGGREGATE OUTFILE - CD AT
  /BREAK = C13
  /AKI TO AK20, AVI TO AV20, All TO AI20, AG1 TO AG20 -
  MEAN(K 1 TO K20, V) TO V20, II TO 120. Gl TO G20)
RECODE SI3 (1=4) (2=5) (3=6)
FILE HANDLE SDAT/NAME = »S.DAT«
AGGREGATE OUTFILE - SDAT
  /BREAK - SI3
  /AKI TO AK20, AVI TO AV20, All TO AI20, AG1 TO AG20 =
  MEAN(K1 TO K20, VI TO V20, 11 TO 120, Gl TO G20)
```

In der ersten DO REPEAT - Schleife werden die ordinalen Dummy-Variablen des Klassifikationsmerkmals AKERNF erzeugt. Die COMPUTE - Anweisung bewirkt zunächst, daß die ordinalen Dummy-Variablen auf 0 gesetzt werden. Durch die IF - Anweisung erhalten sie den Wert 1, sofern der Wert von AKERNF größer/ gleich dem Wert der Variablen WERT ist. Insgesamt werden in der DO REPEAT - Schleife 20 ordinale Dummy-Variablen erzeugt, dadurch ist gewährleistet, daß der maximale Wert von AKERNF erfaßt ist. Dieses Vorgehen wird für die drei anderen Klassifikationsmerkmale wiederholt. Eine Gewichtung ist in diesem Fall nicht erforderlich, da alle Klassifikationsmerkmale dieselbe theoretische Variationsbreite besitzen (vgl. Kapitel 4). Allerdings unterscheiden sich die

tatsächlich empirisch auftretenden Variationsbreiten, was eine Gewichtung rechtfertigen könnte.

Schließlich werden durch die AGGREGATE - Befehle die Verteilungen der berechneten Cluster in den Klassifikationsmerkmalen zwischengespeichert.

Übungsaufgabe **11:** In einer Clusteranalyse wurden u.a. die nominalen Klassifikationsobjekte
REL (Religion) mit den Ausprägungen 1»=röm.kath, 2 = evang., 3 = jüdisch und 9« sonstiges und
FAMST (Familienstand) mit den Ausprägungen 1*ledig, 2 = verh., S^geschieden und 4«verwitwet einbezogen.
Die Daten stehen auf der externen Datei F.DAT. Der Complete Linkage mit der City-Blockmetrik hat eine befriedigende 3-Clusterlösung erbracht.
Schreiben Sie ein SPSS X Programm, in dem die City Blockmetrik zwischen dem Complete - und Baverage-Linkage berechnet wird.

33 Weitere Verfahren

3.3.1 Die Sensitivitätsanalyse

Bei der Sensitivitätsanalyse werden die Klassifikationsmerkmale mit Zufallsfehlern überlagert. Bezeichnen wir allgemein mit X_{il} die empirisch beobachtete Ausprägung des Klassifikationsobjekts i in dem Klassifikationsmerkmal 1, und mit E_{il} einen zufälligen Fehler des Klassifikationsobjektes g in X_l dann wird bei der Sensitivitätsanalyse eine neue Variable Y_l mit

$$Y_{jl} = X_{il} + E_{jl}$$

gebildet.

Für den Zufallsfehler E_{jl} muß eine bestimmte - aus der theoretischen Statistik bekannte - Wahrscheinlichkeitsverteilung angenommen werden, z.B. die Normalverteilung. Die Normalverteilung ist bekanntlich durch zwei Parameter gekennzeichnet: den Erwartungswert μ und die Varianz σ^2 . In Anlehnung an die klassische Testtheorie wird bei der Sensitivitätsanalyse angenommen, daß $\mu = 0$ gilt und daß σ^2 nur eine Funktion des Klassifikationsmerkmals X_l ist. Wir führen deshalb für σ^2 den Index 1 ein.

Für die praktische Durchführung muß nun für σ_l^2 bzw. für σ_l ein bestimmter Wert angenommen werden. Eine andere Möglichkeit besteht darin, σ_l^2 bzw. σ_l systematisch zu variieren. Bei beiden Vorgehensweisen kann σ_i^2 bzw. σ_l nun in Abhängigkeit von

- der beobachteten Varianz s_l^2 von X_l oder von
- der theoretischen Skaleneinheit von X_l,

spezifiziert werden. Im ersten Fall wird man für σ_l^2 einen bestimmten Anteil pan s_l^2 annehmen, z.B. 0.05, 0.10,0.20 usw. Der Wert für σ_l^2 berechnet sich dann nach der Formel:

$$\sigma_l^2 = p^* s_l^2$$

Die zweite Möglichkeit soll anhand unseres Beispiels demonstriert werden. Die Skaleneinheit unserer Klassifikationsmerkmale ist 1.0. Man kann nun $\bar{\sigma}_l^2$ oder σ_l systematisch mit einem vielfachen dieser Skaleneinheit durchvariieren, also z.B. mit den Werten 1.0, 2.0, 3.0 und 4.0.

Technisch läßt sich diese Überlagerung mit Zufallsfehlern mit Hilfe des folgenden SPSSX Programms realisieren:

```
FILE HANDLE AFAMD / NAME - »AFAM.DAT«
GET FILE AFAMD
DO REPEAT ZKERN - ZKERN1 TO ZKERN4/ ZSA •= 1 TO 4
COMPUTE ZKERN - AKERNF+ NORMAUZSA)
END REPEAT
DO REPEAT ZVERW - ZVERW1 TO ZVERW4/ ZSA - 1 TO 4
COMPUTE ZVERW - AVERW+NORMAL(ZSA)
END REPEAT
DO REPEAT ZINW « ZINW1 TO ZINW4/ ZSA * 1 TO 4
COMPUTE ZINW « AINW + NORMAL(ZSA)
END REPEAT
DO REPEAT ZGESIN = ZGESIN1 TO ZGESIN4/ ZSA - 1 TO 4
COMPUTE ZGES1N «= ZGESIN + NORMAL(ZSA)
END REPEAT
CLUSTER AKERNF AVERW A1NW AGESIN
  /MEASURE - BLOCK
  /METHOD - COMPLETED)
  /PRINT - NONE
  /PLOT = NONE
  /SAVE - CLUSTERS)
CLUSTER ZKERN 1 ZVERW 1 Z1NW1 ZGESIN 1
  /MEASURE - BLOCK
  /METHOD - COMPLETE(ZCl)
  /PRINT = NONE
  /PLOT - NONE
  /SAVE = CLUSTERS)
CLUSTER ZKERN2 ZVERW2 ZINW2 ZGESIN2
  /MEASURE = BLOCK
  /METHOD - COMPLETE(ZC2)
  /PRINT - NONE
  /PLOT = NONE
  /SAVE - CLUSTERS)
usw...
```

Die erste DO REPEAT - Schleife bewirkt, daß jedes Klassifikationsobjekt in dem Klassifikationsmerkmal AKERNF mit vier unterschiedlichen normalverteilten Zufallsfehlern überlagert wird. Die Zufallsfehler werden durch den in SPSSX enthaltenen Zufallszahlengenerator NORMAL (σ)

für normalverteilte Zufallszahlen erzeugt. Die mit den Zufallsfehlern überlagerten neuen Klassifikationsmerkmale werden mit ZKERN1, ZKERN2, ZKERN3 und ZKERN4 bezeichnet. Diese sind:

ZKERN1 = AKERNF + zufällige Realisierung einer normal verteilten Zufallsvariablen mit einer Standardabweichung (ZSA) von 1.0.

ZKERN2 = AKERNF + zufällige Realisierung einer normal verteilten Zufallsvariablen mit einer Standardabweichung (ZSA) von 2.0.

ZKERN3 = AKERNF + zufällige Realisierung einer normalverteilten Zufallsvariablen mit einer Standardabweichung (ZSA) von 3.0.

ZKERN4 = AKERNF + zufällige Realisierung einer normal verteilten Zufallsvariablen mit einer Standardabweichung (ZSA) von 4.0.

Dasselbe Vorgehen wird für die verbleibenden Klassifikationsmerkmale AVERW, AINW und AGESIN angewendet. Daran anschließend werden Clusteranalysen gerechnet und die Ergebnisse für weitere Analysen zwischengespeichert. Bei einer Standardabweichung von 1.0 beträgt das symmetrische Lambda zwischen der ursprünglichen und der mit einem Zufallsfehler überlagerten 3-Clusterlösung nur mehr 0.35.

Die Verwendung von normalverteilten Zufallsvariablen hat den Nachteil, daß die mit Zufallsfehlern überlagerten neuen Klassifikationsmerkmale negative Werte annehmen können und daß die Werte bis ± ∞ variieren können. Diese extremen Werte treten empirisch zwar nicht auf, dennoch können zufällig hohe oder niedere Werte das Ergebnis vollkommen zerstören. Zur Vermeidung dieses Effektes können anstelle von normalverteilten Zufallsvariablen gleichverteilte Zufallsvariablen erzeugt werden. In diesem Fall kann z.B. angenommen werden, daß ein bestimmter Prozentsatz jedes Klassifikationsmerkmals fehlerbehaftet ist, z.B. mit 5%, 10% oder 20%, und daß die Fehler maximal zwischen ± einer bestimmten Anzahl liegen, z.B. zwischen ± 1 oder zwischen ± 2 usw..

In dem nachfolgenden SPSS-X Programm wurden beispielsweise für unsere Daten ein Fehleranteil von 10 % und eine maximale Fehlervariation von ± 2 angenommen.

```
FILE HANDLE AFAMD / NAME = »AFAM.DAT«
GET FILE AFAMD
COMPUTE FEHLER - UNIFORM(IOO)
COMPUTE FEHLK - 0
IF (FEHLER LE 10 ) FEHLK - RND(UNIFORM(4) + 0.5)
RECODE FEHLK (0 = 0) (1 = - 2) (2 - - 1) (3=1) (4 = 2)
COMPUTE ZKERN - AKERNF + FEHLK
COMPUTE FEHLER - UNIFORM(100)
```

```
COMPUTE FEHLK - 0
IF (FEHLER LE 10 ) FEHLK - RND(UNIFORM(4) + 0.5)
RECODE FEHLK (0-0) (1= - 2) (2- - 1) (3=1) (4 = 2)
COMPUTE ZVERW - AVERW + FEHLK
COMPUTE FEHLER « UNIFORM(IOO)
COMPUTE FEHLK - 0
IF (FEHLER LE 10 ) FEHLK = RND(UNIFORM(4) + 0.5)
RECODE FEHLK (0=0) (1= -2) (2= - 1) (3=1) (4 = 2)
COMPUTE ZINW = A1NW + FEHLK
usw..
```

Durch die Anweisung UNIFORM(IOO) wird eine im Intervall (0,100) gleichverteilte Zufallsvariable erzeugt. Diese wird durch den COMPUTE - Befehl der Variablen FEHLER zugewiesen. Durch die Anweisung COMPUTE FEHLK = 0 wird zunächst der Fehler FEHLK auf 0 gesetzt. Die nachfolgende IF - Anweisung bewirkt, daß nur in 10% der Fälle, also wenn die Variable FEHLER kleiner/gleich 10 ist, ein zufälliger Fehler FEHLK berechnet wird. Diese Variable nimmt Werte zwischen 1 und 4 an, da durch die Anweisung RND(UNIFORM(4)+ 0.5) zunächst eine gleichverteilte Zufallszahl zwischen 0 und 4 erzeugt wird. Dieser Zahl wird anschließend der Wert 0.5 addiert und gerundet. Die Werte 1,2,3 und 4 treten dadurch mit gleicher Wahrscheinlichkeit auf. In der nachfolgenden RECODE - Anweisung wird der Fehler FEHLK auf die Werte +2, +1, -1 und - 2 transformiert. Durch die auf den RECODE-Befehl folgende COMPUTE - Anweisung wird das Klassifikationsmerkmal AKERNF mit dem zufälligen Fehler FEHLK, sofern dieser auftritt, überlagert.

Diese Schritte werden nun für die verbleibenden Klassifikationsmerkmale AVERW, AINW und AGESIN wiederholt. Zwar können auch bei diesem Vorgehen negative Werte auftreten, die sich aber in bescheidenen Grenzen halten.

Für diese Eingabeparameter ergibt sich ein symmetrisches Lambda von 0.50 zwischen der ursprünglichen und der mit einem Meßfehler überlagerten 3-Clusterlösung.

Das Vorgehen mit gleichverteilten Zufallszahlen hat den Vorteil, daß es sich unmittelbar auf nominale Variablen übertragen läßt.

Bis jetzt liegen noch wenige Ergebnisse vor, wie sich zufällige Fehler auf die Stabilität einer Clustcrlösung auswirken. Erste Simulationsexperimente (Bacher 1987) zeigen, daß unter den untersuchten Clusteranalyseverfahren (Baverage - , Complete - und Single Linkage) das Complete-Linkageverfahren gegenüber zufälligen Fehlern sehr robust ist.

Tabelle 3.3-1:

Durchschnittlicher Anteil von Fehlklassifikationen für unterschiedliche hierarchische Clusterverfahren in Abhängigkeit von zufälligen Fehlern.

	Anteil zufälliger Meßfehler an der »wahren« Varianz				
Verfahren	20%	36%	50%	64%	75%
Single	53.3%	71.1%	71.3%	71.3%	71.3%
Baverage	14.6%	37.3%	47.7%	54.7%	60.1%
Complete	4.4%	25.4%	34.4%	48.0%	54.1%

Bezüglich der untersuchten Distanzmaße (City-Blockmetrik und euklidische Metrik) führte die euklidische Metrik zu geringeren Fehlklassifikationen der Klassifikationsobjekte als die City-Blockmetrik.

Tabelle 3.3-2:

Durchschnittlicher Anteil von Fehlklassifikationen für die City-Blockmetrik und die euklidische Metrik beim CompleteLinkage

	Anteil zufälliger Meßfehler an der »wahren« Varianz				
Metrik	20%	36%	50%	64%	75%
Block	0.0%	17.0%	39.0%	53.0%	57.0%
Euklid	1.0%	13.0%	33.0%	46.0%	56.0%

Übungsaufgabe 12: Gegeben sei die Datei F.DAT der Übungsaufgabe 11. Die Klassifikationsmerkmale REL und FAMST sollen mit einem zufälligen Meßfehler überlagert werden. Bezüglich der zufälligen Meßfehler werden folgende Annahmen getroffen:

- Der zufällige Meßfehler tritt bei jeder Ausprägung mit einer Wahrscheinlichkeit von 0.1 auf.
- Die Fehler verteilen sich zufällig auf die anderen Ausprägungen.

Schreiben Sie das entsprechende SPSS-X Programm, das diese Aufgabe löst.

3.3.2 Die Diskriminanzanalyse

In der Literatur wird sehr häufig die Diskriminanzanalyse zur Überprüfung einer Clusterlösung eingesetzt. Sie ist wie die Clusteranalyse ein multivariates statistisches Verfahren, das aber im Unterschied zur hierarchischen Clusteranalyse voraussetzt, daß die Anzahl der Cluster und zumindest für einen Teil der Klassifikationsobjekte deren Clusterzugehörigkeit bekannt sind. Anstelle von Clustern wird in der Literatur zur Diskriminanzanalyse die Bezeichnung Gruppen verwendet.

Die Diskriminanzanalyse versucht, aus den Klassifikationsmerkmalen sogenannte Diskriminanzfunktionen zu berechnen, die die Gruppen am besten trennen. Diese Diskriminanzfunktionen können in einem weiteren Schritt der Analyse zur Berechnung der a posteriori Zuordnungswahrscheinlichkeiten der Klassifikationsobjekte zu den Clustern verwendet werden. Diese geben an, mit welcher Wahrscheinlichkeit ein Klassifikationsobjekt mit einem bestimmten Wert in den Klassifikationsmerkmalen einem der bekannten Cluster angehört. Dadurch werden Fehlzuordnungen sichtbar, wenn z.B. die Zuordnungswahrscheinlichkeit eines Klassifikationsobjektes i für das Cluster k 1.0 beträgt und das Klassifikationsobjekt ursprünglich dem Cluster l ($l \neq k$) zugerechnet wurde.

Die Diskriminanzfunktionen und damit die a posteriori Wahrscheinlichkeiten können aber auch für Klassifikationsobjekte berechnet werden, für die zwar die Ausprägungen in den Klassifikationsmerkmalen, nicht aber die Clusterzugehörigkeit bekannt ist. Damit ist eine Zuordnung dieser Klassifikationsobjekte möglich, aber keine Überprüfung der Zuordnung.

Zusammenfassend können in bezug auf Klassifikationsaufgaben zwei Anwendungsfälle der Diskriminanzanalyse unterschieden werden:

1. die **Stabilitätsprüfung** einer Clusteranalyse und
2. die **Zuordnung von Klassifikationsobjekten** bei großen Datensätzen

Da bei den hierarchischen Clusteranalyseverfahren eine Distanzmatrix zwischen den Klassifikationsobjekten berechnet werden muß, ist ihre Anwendung auf kleine und mittlere Datensätze begrenzt. Bei einem großen Datensatz kann deshalb so vorgegangen werden:

1. Wähle zufällig eine Unterstichprobe aus.
2. Führe für diese Unterstichprobe eine Clusteranalyse aus.
3. Uberprüfe die Stabilität dieser Clusterlösung, z.B. mit Hilfe einer Diskriminanzanalyse, und führe eventuell Modifikationen durch.
4. Liegt eine stabile Clusterlösung vor, berechne die Diskriminanzfunktionen und Zuordnungswahrscheinlichkeiten der nicht in der Unterstichprobe enthaltenen Klassifikationsobjekte und nimm eine Zuordnung zu den Clustern vor.

In diesem Abschnitt wird nur die Anwendung der Diskriminanzanalyse zur Stabilitätsprüfung behandelt und dabei nur die lineare Diskriminanzanalyse dargestellt. (Eine ausführliche Darstellung der linearen Diskriminanzanalyse und von anderen Modellen der Diskriminanzanalyse geben z.B. Andrews 1972, Fahrmeir u.a. 1984 oder Klecka 1984).

Das Modell der linearen Diskriminanzanalyse:

Gegeben seien p quantitative Klassifikationsmerkmale $X_1, X_2, ..,$ eine Clusterlösung mit c Clustern und n Klassifikationsobjekte. Für diese Klassi-

fikationsobjekte sind die Zugehörigkeit zu einem Cluster j und die Ausprägungen in den Klassifikationsmerkmalen bekannt. Mit X_{ijk} soll die Ausprägung des Klassifikationsobjektes k im j - ten Cluster in dem Klassifikationsmerkmal X_i bezeichnet werden.

Gesucht werden nun Linearkombinationen $L_1, L_2, \ldots$ der ursprünglichen Klassifikationsmerkmale $X_1, X_2, \ldots$ der Art:

$$\begin{aligned} L_{1jk} &= d_{11}X_{1jk} + d_{12}X_{2jk} + \ldots + d_{1p}X_{pjk} \\ L_{2jk} &= d_{21}X_{1jk} + d_{22}X_{2jk} + \ldots + d_{2p}X_{pjk} \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \end{aligned}$$

Diese Linearkombinationensollen die Cluster am »besten« trennen. Eine beste Trennung soll dann vorliegen, wenn das Verhältnis der Streuung zwischen den Clustern (SS_B) zu dem der Streuung innerhalb des Clusters (SS_W) maximal ist. Wenn also gilt:

$$\frac{SS_B}{SS_W} = \max!$$

Die Streuung zwischen den Clustern ist ein Maß für die Heterogenität der Cluster. Je größer SS_B ist, umso »besser« werden die Mittelwerte der Cluster in den Klassifikationsobjekten voneinander getrennt. Eine Maximierung von SS_B ist also ganz offensichtlich sinnvoll, da dadurch eine »beste« Trennung der Cluster erreicht wird, sofern die Mittelwerte der Cluster als repräsentative Merkmale der Cluster aufgefaßt werden (9). Die Streuung innerhalb der Cluster (SS_W) dagegen, ist ein Maß für die Heterogenität innerhalb der Cluster. Sie sollte möglichst klein sein. Mathematisch präzisiert bedeutet dieses »möglichst klein«, daß SS_W minimiert oder äquivalent dazu der inverse Wert $1/SS_W$ maximiert werden soll.

Die Linearkombinationen $L_1, L_2, \ldots$ werden schrittweise bestimmt. Zunächst wird eine Linearkombination L_1 gesucht, die das obige Kriterium maximiert, die also die Cluster am besten trennt. Im nächsten Schritt wird eine zweite Linearkombination L_2 gesucht, die die Cluster am zweit besten trennt, dann eine Linearkombination L_3, die die Cluster am dritt besten trennt, usw.. Insgesamt gibt es maximal q Linearkombinationen, wobei $q = c - 1$ ist, wenn die Anzahl p der Klassifikationsmerkmale größer/gleich c - 1 ist (c ist die Anzahl der Cluster). In den anderen Fällen ist $q = p$. In unserem Beispiel der familialen Haushaltsdaten mit einer 3-Clusterlösung gibt es also zwei Diskriminanzfunktionen.

Die Erklärungskraft jeder Linearkombination wird durch ihren Eigenwert gemessen. Allgemein gilt, daß die Summe aller Eigenwerte gleich dem Wert des Maximierungskriteriums ist. Division jedes Eigenwertes mit dieser Gesamtsumme, ergibt deshalb den Erklärungsanteil jeder Linear-

kombination zur Trennung der Cluster. Dieser Anteil gibt den Prozentsatz an, mit welchem die Cluster durch die entsprechende Linearkombination getrennt werden.

In unserem Beispiel sind diese Eigenwerte und Erklärungsanteile:

Eigenwert der 1. Linearkombination	= 2.15 (62.5%)
Eigenwert der 2. Linearkombination	= 1.29 (37.5%)
Gesamt	= 3.44 (100 %)

Die erste Linearkombination leistet also einen Beitrag von ungefähr 63% zur Trennung der Cluster, die zweite einen Beitrag von ungefähr 38%.

In SPSSX kann zusätzlich ein Test für die Signifikanz der Linearkombinationen durchgeführt werden. Dieser Test ist hierarchisch aufgebaut. Zunächst wird geprüft, ob die 1.,2.,3.,.. und q - te Linearkombination signifikant sind. Im zweiten Schritt, ob die 2.,3.,.. und q - te Linearkombination signifikant sind, im dritten, ob die 3.,..und q - te Linearkombination signifikant sind, usw... In jedem Schritt wird als Teststatistik das sogenannte Wilks Lambda berechnet. Für die einzelnen Schritte ist dieses wie folgt definiert:

Schritt 0: Signifikanz der 1.,2.,3,... und q - ten Linearkombination

$$\text{Wilks Lambda} = \frac{1}{1+v_1} * \frac{1}{1+v_2} * \ldots * \frac{1}{1+v_q}$$

Schritt 1: Siknifikanz der 2.,3.,.. und q - ten Linearkombination

$$\text{Wilks Lambda} = \frac{1}{1+v_2} * \ldots * \frac{1}{1+v_q}$$

.
.
.

Schritt q - 1: Signifikanz der q - ten Linearkombination

$$\text{Wilks Lambda} = \frac{1}{1+v_2}$$

Das Wilks Lambda läßt sich in eine Chi - quadratverteilte Zufallsgröße mit (p - k)*(c - k - 1) überführen. Der Test des Schrittes k gibt also an, ob alle nach der k - ten Linearkombination auftretenden Linearkombinationen signifikant sind oder nicht.

Die so berechneten Linearkombinationen werden als kanonische Diskriminanzfunktionen bezeichnet. Kanonisch deshalb, da sie zusätzlich die

Eigenschaften der paarweisen Unabhängigkeit besitzen. Sie bilden die Basis für die Berechnung folgender Maßzahlen:

1. der standardisierten Diskriminanzkoeffizienten,
2. der Korrelationen der Klassifikationsmerkmale mit den kanonischen Diskriminanzfunktionen (pooled within - groups correlations) und
3. der unstandardisierten Diskriminanzkoeffizienten.

Die beiden ersten Maßzahlen sind Hilfsgrößen für die Interpretation der Diskriminanzfunktionen. Die **standardisierten Diskriminanzkoeffizienten** können wie **standardisierte Regressionskoeffizienten** interpretiert werden. Sie messen also den **direkten Beitrag** jedes Klassifikationsmerkmals zur Erklärung der Diskriminanzfunktionen, unter der Annahme, daß alle Klassifikationsmerkmale in der gleichen Skaleneinheit gemessen wurden. Das bedeutet aber zugleich, daß die standardisierten Diskriminanzkoeffizienten von zwei stark miteinander korrelierten Klassifikationsmerkmalen niedrige Werte annehmen, da sich die beiden Merkmale gegenseitig Erklärungskraft wegnehmen. Aus diesem Grund werden zusätzlich Korrelationen zwischen den Klassifikationsmerkmalen und den kanonischen Diskriminanzfunktionen berechnet, bei denen diese gegenseitigen Abhängigkeiten nicht berücksichtigt werden.

In unserem Beispiel nehmen diese beiden Maßzahlen folgende Werte an (vgl. Tabelle 3.3-3).

Aus der Tabelle ist ersichtlich, daß die erste Diskriminanzfunktion von dem Klassifikationsmerkmal AVERW gebildet wird, die zweite von der Anzahl der Mitglieder der Kernfamilie (AKERNF). Die erste Diskriminanzfunktion könnte deshalb als »reine Verwandtschaft ohne Kernfamilie« interpretiert werden, die zweite dagegen als »Kernfamilie ohne Verwandtschaft«. Dieser Interpretation können aber die Einwände, wie sie bei der Interpretation der Ergebnisse des CompleteLinkage gemacht wurden (vgl. Abschnitt 2.5) entgegengehalten werden. (Um eine inhaltlich bessere Interpretation der Diskriminanzfunktionen zu erhalten, kann in SPSS-X eine orthogonale Rotation der Diskriminanzfunktionen durchgeführt werden.)

Die nichtstandardisierten Diskriminanzkoeffizienten werden dagegen für die Berechnung der Zuordnungswahrscheinlichkeiten benötigt. Diese sind in der Tabelle 3.3-4 dargestellt.

Tabelle 3.3-3:

Standardisierte Diskriminanzkoeffizienten und Korrelationen zwischen den Klassifikationsmerkmalen und den kanonischen Diskriminanzfunktionen für die 3-Clusterlösung des Complete-Linkage der familialen Haushaltsdaten

Klassifikationsmerkmale	Standardisierte kanonische Diskr.funktion 1	Diskriminanzkoeffizienten kanonische Diskr.funktion 2
AKERNF	.15	1.01
AVERW	1.04	.07
AINW	-.20	-.00
AGESIN	-.12	-.32
Klassifikationsmerkmale	**Korrelationen zwischen Kl.merkmalen und kanonische Diskr.funktion 1**	**kanonische Diskr.funktion 2**
AKERNF	-.05	.94
AVERW	.95	-.12
AINW	-.06	-.14
AGESIN	-.06	-.17

Tabelle 3.3-4:

Unstandardisierte Diskriminanzkoeffizenten

	Diskr.funktion 1	Diskr.funktion 2
AKERNF	.00	.64
AVERW	.77	.06
AINW	-.62	.03
AGESIN	-.14	.35
Konstante	-1.71	-2.87

Auf ihrer Grundlage werden zunächst die sogenannten Diskriminanzscores für jedes Klassifikationsobjekt und für die Cluster berechnet. Die Formel zur Berechnung ist:

$$D_{ijk} = a_i + b_{i1}X_{1jk} + b_{i2}X_{2jk} + ... + b_{ip}X_{pjk}$$

mit D_{ijk} = Diskriminanzscore des Klassifikationsobjektes j des Clusters k in der i - ten kanonischen Diskriminanzfunktion

a_i = konstanter Faktor der i - ten Diskriminanzfunktion

b_{ii} = unstandardisierter Diskriminanzkoeffizient des Klassifikationsmerkmals X_i in der i - ten Diskriminanzfunktion

Für die Gruppenzentroide (Clustermittelwerte) erhält man die in der Tabelle 3.3-5 dargestellten Diskriminanzscores. Das dritte Cluster ist durch einen hohen Wert auf der ersten Diskriminanzfunktion, der Verwandtschaftsdimension, gekennzeichnet, das zweite durch einen hohen Wert auf der zweiten Diskriminanzfunktion. Die Lage der Zentroiden kann auch graphisch dargestellt werden (vgl. Abbildung 3.3-1).

Tabelle 3.3-5:

Diskriminanzscores der Gruppenzentroide

Gruppe (Cluster)	Diskr.funktion 1	Diskr.funktion 2
1	-.88	-.64
2	-.30	2.23
3	2.80	-.33

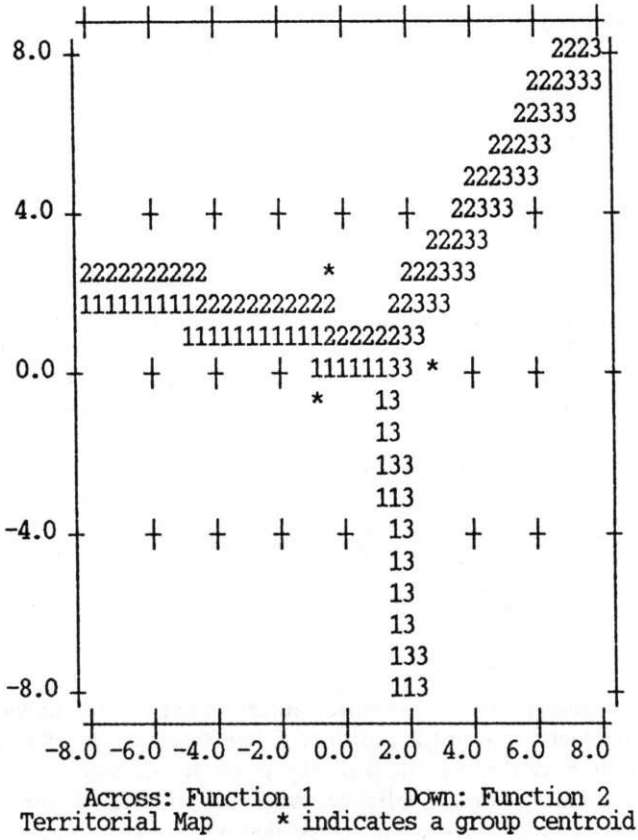
Die Abbildung 3.3-1 enthält zusätzlich die Trennlinien zwischen den Clustern (Gruppen). Alle Klassifikationsobjekte, die in der durch die Linie »1« begrenzten Fläche liegen, gehören dem ersten Cluster (der ersten Gruppe) an.

Für die Berechnung der Zuordnungswahrscheinlichkeiten wird noch zusätzlich die VarianzKovarianzmatrix innerhalb jedes Clusters benötigt. Diese drei Größen, die Diskriminanzscores der Klassifikationsobjekte und die der Cluster (Gruppen)zentroide) und die Varianz Kovarianzmatrizen innerhalb jedes Clusters ermöglichen nun eine Berechnung folgender Wahrscheinlichkeiten:

- Die bedingte Wahrscheinlichkeit $P(X_{1jk}, \ldots, X_{pjk}/k)$ ist Element von j), daß die empirisch beobachteten Ausprägungen des Klassifikationsobjektes k in den Klassifikationsmerkmalen auftreten unter der Voraussetzung, daß k dem Cluster j angehört. In SPSS-X wird diese Wahrscheinlichkeit mit P(D/G) (Probability of data given group) ausgegeben.
- Die bedingte Wahrscheinlichkeit (Zuordnungswahrscheinlichkeit) $P(k \text{ ist Element von } j / X_{1jk}, \ldots, X_{pjk})$. daß ein Klassifikationsobjekt k dem Cluster j angehört unter der Voraussetzung, daß die empirisch beobachteten Ausprägungen in den Klassifikationsmerkmalen gege-

Abbildung 3.3-1:

Graphische Darstellung der Gruppenzentroide und Trennflächen



ben sind. Diese Wahrscheinlichkeit wird in SPSS-X als P(**G**/D) (Probability of group given data) bezeichnet.

Jedes Klassifikationsobjekt wird nun dem Cluster zugeordnet, für das P(G/D) maximal ist.

Die nachfolgende Tabelle enthält diese Wahrscheinlichkeiten und die dazugehörenden Diskriminanzscores für die ersten 15 Klassifikationsobjekte.

Tabelle 3.3-6:

Diskriminanzscores und Zuordnungswahrscheinlichkeiten für die ersten 15 Haushalte

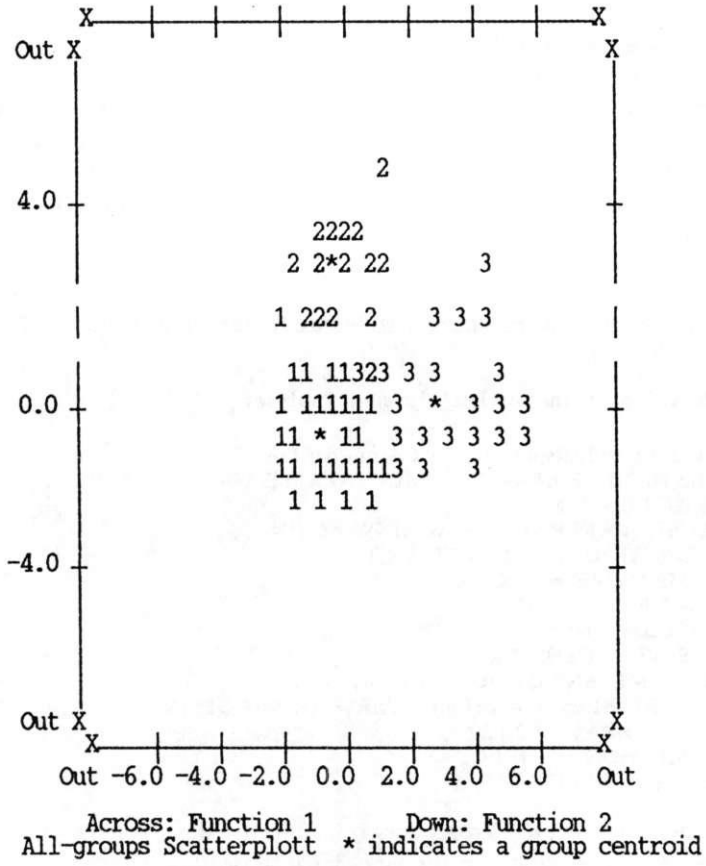
Case	Actual Group		Highest Group	Probability P(D/G)	P(G/D)	2nd Highest Group	P(G/D)	Discriminant Scores	
1	1		1	.44	1.00	2	.00	-2.04	-1.18
2	1		1	.51	.97	3	.03	0.27	-0.75
3	1		1	.59	.94	2	.06	1.22	0.33
4	1		1	.59	.98	3	.02	0.12	-0.83
5	1		1	.82	1.00	2	.00	-1.42	-0.95
6	2		2	.89	.96	1	.04	-0.77	2.34
7	1		1	.85	.99	2	.01	-1.32	-0.31
8	1		1	.21	1.00	2	.00	-1.61	-2.23
9	1		1	.49	1.00	2	.00	-2.04	-0.92
10	2		2	.66	.74	1	.26	-1.03	1.61
11	1		1	.46	.91	2	.08	0.84	0.16
12	2		2	.40	.57	1	.40	0.42	1.09
13	3	**	1	.27	.80	2	.17	0.32	0.45
14	1		1	.42	.82	2	.18	-0.52	0.62
15	1		1	.82	.99	2	.02	-1.42	-0.95

Das erste Klassifikationsobjekt wurde aufgrund der Clusteranalyse dem ersten Cluster (actual group) zugeordnet. Diese Gruppe besitzt auch die größte Zuordnungswahrscheinlichkeit für das erste Cluster (P(G/D) = 1.00), d.h., mit einer Wahrscheinlichkeit von 100% tritt das Cluster 1 auf, wenn die Ausprägungen des ersten Klassifikationsobjektes in den Klassifikationsmerkmalen vorliegen. Der Wert von P(D/G) = 0.44 besagt dagegen, daß diese Ausprägungen aber für das erste Cluster als solches weniger typisch sind und nur mit einer Wahrscheinlichkeit von 44% auftreten. Die zweit höchste Zuordnungswahrscheinlichkeit zu einem Cluster liegt schon bei 0%. Darüber hinaus besitzt das erste Klassifikationsobjekt einen schwach negativen Wert in der ersten kanonischen Diskriminanzfunktion und einen sehr starken negativen Wert in der zweiten. Eine Fehlzuordnung tritt zum ersten Mal beim 13. Klassifikationsobjekt auf. Auf der Grundlage der Diskriminanzanalyse wird es dem 1. Cluster zugeordnet, während es bei der Clusteranalyse dem 3. Cluster zugerechnet wurde.

Die Diskriminanzscores können für eine graphische Darstellung verwendet werden (vgl. Abbildung 3.3-2). Die Mittelwerte (Zentroide) der Cluster sind durch einen Stern angedeutet. Besitzen mehrere Klassifikationsobjekte dieselben Diskriminanzscores, werden sie nur einmal dargestellt.

Abbildung 3.3-2:

Graphische Darstellung der Klassifikationsobjekte und Cluster auf den Diskriminanzfunktionen



Die richtigen und fehlerhaften Zuordnungen können zusammenfassend in einer Kreuztabelle dargestellt werden (vgl. Tabelle 3.3-7).

Tabelle 3.3-7:

Anzahl richtiger und fehlerhafter Zuordnungen bei der Diskriminanzanalyse

Zuordnung der Cluster-analyse	Zuordnungen bei der Diskriminanzanalyse		
	Cluster 1	Cluster 2	Cluster 3
Cluster 1	92	2	0
Cluster 2	1	31	0
Cluster 3	4	0'	29

Insgesamt werden durch die Diskriminanzanalyse 95,6% der Klassifikationsobjekte richtig zugeordnet. Dieser Befund könnte durchaus als zufriedenstellend aufgefaßt werden. Er steht aber in deutlichem Widerspruch zu den bisherigen Ergebnissen, insbeonders zu den Ergebnissen der Sensitivitätsanalyse und des Vergleichs von clusteranalytischen Verfahren.

An dieser Stelle kann deshalb nur noch einmal die Empfehlung ausgesprochen werden, mehrere Ansätze der Stabilitätsprüfung zu verfolgen.

Abschließend sollen noch das SPSS·X Programm, das zur Berechnung der in diesem Abschnitt erzielten Ergebnisse führt, dargestellt und die Annahmen der Diskriminanzanalyse zusammengefaßt werden (SPSS Inc. 1986: 688-712).

SPSS·X Programm zur Diskriminanzanalyse:

```
TITLE »Stabilitätsprüfung der CA mit der DA«
FILE HANDLE AFAMD/NAME = »AFAM.DAT«
GET FILE AFAMD
CLUSTER AKERNF AVERW AINW AGSIN
  /METHOD - COMPLETE(Cl)
  /MEASURE - BLOCK
  /PRINT - NONE
  /PLOT = NONE
  /SAVE = CLUSTER (3)
D1SCR1MINANTGROUPS - C130.3)
  /VARIABLES = AKERNF AVERW AINW AGESIN
  /PRIORS - .59 .20 .21
  /METHOD - DIRECT
STAT1STICS 10 13 14 15 16
```

In der CLUSTER - Prozedur wird zunächst die Clusterzugehörigkeit der Klassifikationsobjekte für eine 3-Clusterlösung berechnet und zwischengespeichert. Durch den Befehl D1SCRIMINANT wird die SPSS·X Prozedur D1SCR1MINANT aufgerufen. Daran anschließend wird durch

GROUPS = Cl 3(1,3) die zwischengespeicherte 3-Clusterlösung als Gruppierungsvariable definiert. Allgemein müssen bei der Definition der Gruppierungsvariablen deren Unter - und Obergrenze in den Klammern angegeben werden. Nach der Definition der Gruppierungsvariablen werden die Variablen, die in die Diskriminanzanalyse einbezogen werden sollen, mit VARIABLES = Liste der Variablen definiert. In unserem Beispiel wird diese Variablenliste von den Klassifikationsmerkmalen AKERNF, AVERW, AINW und AGSIN gebildet. Durch die Anweisung PRIORS = .566 .302 .132 werden die a priori Anteile der Gruppengrößen festgelegt. Sie entsprechen in unserem Fall den Größen der Cluster. In SPSS-X müssen diese a priori Wahrscheinlichkeiten nicht eingegeben werden. In diesem Fall entfällt die PRIORS - Anweisung und es werden gleich große Gruppen angenommen.

Die Eingabe von a priori Wahrscheinlichkeiten setzt die Kenntnis der »wahren« Gruppengrößen voraus. Sie ist also nur in jenen Fällen gerechtfertigt, wo die Diskriminanzanalyse zur Überprüfung einer Klassifikation eingesetzt wird und die Ausgangslösung als »wahre« Lösung akzeptiert wird. In der Literatur, insbesondere in der zur medizinischen oder psychiatrischen Diagnose, wird auch noch empfohlen, a priori Wahrscheinlichkeiten vorzugeben, wenn eine kleine Gruppe (Gruppe der Symptomträger) existiert, und eine falsche Zuordnung zu dieser mit einer größeren Gefahr verbundens ist als eine fälschliche Zuordnung zu einer anderen, größeren Gruppe (Gruppe der Gesunden).

Durch die Anweisung METHOD wird das Verfahren zur Berechnung der kanonischen Diskriminanzfunktionen festgelegt. Die Anweisung METHOD = DIRECT bedeutet, daß alle Analysevariablen simultan in die Berechnung der kanonischen Diskriminanzfunktionen einbezogen werden. Daneben bestehen noch Möglichkeiten einer schrittweisen Variablenselektion zur Berechnung der kanonischen Diskriminanzfunktionen.

Durch die STATISTICS - Anweisungen werden die zu berechnenden Statistiken festgelegt, die Zahlen bedeuten dabei:

10 = Plot der Trennlinien
15 = Plot der Lage der Klassifikationsobjekte und der Gruppen auf den Diskriminanzfunktionen
16 = Getrennte Plots für jede Gruppe
13 = Darstellung der Zuordnungen in einer Kreuztabelle
14 = Ausgabe der Zuordnungswahrscheinlichkeiten und der Diskriminanzscores

Die Annahmen der Diskriminanzanalyse:

(1) Es wird ein lineares Modell für die kanonischen Diskriminanzfunktionen angenommen.

(2) Die Diskriminanzkoeffizienten sind für alle Gruppen (Cluster) und Klassifikationsobjekte gleich.

(3) Es können nur quantitative Klassifikationsmerkmale einbezogen werden. (Nominale oder ordinale Klassifikationsmerkmale müssen zuvor in Dummy-Variablen aufgelöst werden.)

(4) Das hier dargestellte Verfahren geht auf Fisher zurück. Bei diesem Verfahren ist für die Berechnung der Diskriminanzscores keinesfalls **eine** multivariate Normal**Verteilung** oder die Gleichheit der Varianz-Kovarianzmatrizen innerhalb den Gruppen erforderlich (vgl. Fahrmeir, Häußler *Sc* Tutz 1984: 316 - 323). Die Gleichheit der Varianz-Kovarianzmatrizen innerhalb den Gruppen und die Annahme einer multivariaten Normal Verteilung geht dagegen in die statistischen Signifikanztests **für** die Anzahl der Diskriminanzfunktionen und in eine automatische Variablenselektion ein, sofern diese gewählt wird.

(5) Die Zuordnungswahrscheinlichkeiten P(D/G) werden auf der Grundlage **von** Mahalanobisdistanzen und den a priori Wahrscheinlichkeiten berechnet. Da die a priori Wahrscheinlichkeiten nur in die Berechnung der Zuordnungswahrscheinlichkeiten eingehen, ergeben sich für die Eigenwerte, die standardisierten und nichtstandardisierten Diskriminanzkoeffizienten, sowie für die Korrelationen der Klassifikationsmerkmale mit den Diskriminanzfunktionen bei Verwendung derselben Schätzmethode immer identische Werte, unabhängig davon, welche a priori Wahrscheinlichkeiten gewählt werden. Die Zuordnungswahrscheinlichkeiten ändern sich dagegen.

(6) In die Berechnung der Wahrscheinlichkeiten P(D/G) geht die Annähme einer multivariaten Normalverteilung ein.

(7) Bei einer Verletzung der Annahme der Gleichheit der Varianz-Kovarianzmatrix und der multivariaten Normalverteilung sind die Test-**Statistiken** zunächst verzerrt. Darüber hinaus liegt keine optimale Entscheidungsregel (Zuordnungsregel) mehr vor, bei der die Gesamtfehlerrate falscher Zuordnungen ein Minimum ist. Allerdings sind die Verzerrungen bei nicht zu stark unterschiedlichen Varianz-Kovarianzmatrizen innerhalb der Gruppen nur geringfügig, während die Verletzung der Normalverteilung - außer bei Dummy-Variablen - einen stärkeren Effekt hat (vgl. Fahrmeir, Häußler *Sc* Tutz 1984).

Übungsaufgabe 13:

a) Interpretieren Sie die Zuordnungswahrscheinlichkeiten und Diskriminanzscores der Klassifikationsobjekte 2-5 der Tabelle 3.2-5.

b) Nominale und ordinale Klassifikationsmerkmale können durch die Auflösung in Dummy-Variablen in die Analyse einbezogen werden. Verwenden Sie die Datei F.DAT der Übungsaufgabe 11 und schreiben Sie das entsprechende SPSS X Programm, das eine 4-Clusterlösung des Single-Linkage bei Verwendung der City-Blockmetrik überprüft.

c) Wieviele Diskriminanzfunktionen können in b) berechnet werden? Begründen Sie Ihre Antwort!

Historical Social Research, Vol. 14 — 1989 — No. 2, 6-167

4. Fehlende Werte, Meßfehler und die Transformation von Klassifikationsmerkmalen

4.1 Das Problem fehlender Werte

In vielen Anwendungsfällen wird die Klassifikationsdatenmatrix fehlende Werte aufweisen: In den familialen Haushaltsdaten wäre das z.B. der Fall, wenn für einige Haushalte Angaben über die Kernfamilie, die Verwandten, die Inwohner und/oder das Gesinde fehlten. In der Tabelle 4.1-1 wurden diese fehlenden Werte künstlich erzeugt und sind mit - 9 gekennzeichnet. Im zweiten Haushalt fehlen Angaben über die Anzahl der Inwohner, im dritten über das Gesinde und im vierten über die Anzahl der Verwandten.

Tabelle 4.1-1:

Die Ausprägungen der ersten fünf Klassifikationsobjekte der gewünschten Klassifikationsdatenmatrix

Klassifikations-Objekte	Klassifikationsmerkmale AKERNF	AVERW	AINW	AGESIN
HH 1	1.00	0.00	0.00	3.00
HH 2	3.00	3.00	-9	0.00
HH 3	5.00	-9	0.00	0.00
HH 4	3.00	2.00	0.00	-9
HH 5	3.00	0.00	0.00	0.00

4.1.1 Fallweise Elimination

In der SPSS X Prozedur CLUSTER werden fehlende Werte sehr rigoros gehandhabt. Alle Klassifikationsobjekte (Fälle) mit einem oder mehreren fehlenden Werten werden aus der Analyse eliminiert. Dieses Vorgehen wird als fallweise Elimination bezeichnet. In der Tabelle 4.1-1 würde diese zur Elimination der Hauhalte 2, 3 und 4 führen.

4.1.2 Paarweises Ausscheiden

Die **fallweise Elimination** kann zu einer **Verzerrung** der **Repräsentativität der Klassifikationsobjekte** führen. Eine einfache alternative Strategie ist das - aus der Korrelationsrechnung bekannte - paarweise Ausscheiden. In die Berechnung der Un - oder Ähnlichkeit zwischen zwei Klassifikationsobjekten i und j werden nur jene Klassifikationsmerkmale einbezogen, in denen beide Klassifikationsobjekte gültige (nicht fehlende) Werte besitzen. In die Berechnung der Un - oder Ähnlichkeit zwischen dem 1. und 2. Haushalt würden beispielsweise die Klassifikationsmerkmale AKERNF, AVERW und AGESIN eingehen, da der 2. Haushalt in AINW einen fehlenden Wert besitzt. Für die Berechnung der Un - oder Ähnlichkeit zwischen dem 2. und 3. Haushalt würden die Klassifikationsmerkmale AVERW und AINW eliminiert, da der 2. Haushalt einen fehlenden Wert in AINW besitzt und der 3. Haushalt in AVERW aufweist. Erforderlich ist bei diesem Vorgehen eine Reskalierung, so daß jedem Un - oder Ähnlichkeitswert die gleiche Anzahl von Klassifikationsmerkmalen zugrundeliegt.

Die allgemeine Formel zur Reskalierung bei der Verwendung der Minkowski-Metrik ist:

$$\text{POWER}(i,j/p,r) = \left|(m/g(i,j))\Sigma g(i,j,l)\,[\text{ABS}(X_{il} - X_{jl})]^p\right|^{1/r}$$

mit

X_{il} = Ausprägung des Klassifikationsobjektes 1 in dem Klassifikationsmerkmal 1

$g(i,j,l)$ = 1 wenn i und j in dem Klassifikations - merkmal gültige Werte besitzen

0 sonst

$g(i,j) = \Sigma g(i,j,l)$ = Anzahl der gültigen Werte der 1 Klassifikationsobjekte i und j in den Klassifikationsmerkmalen

m = Anzahl der Klassifikationsmerkmale

p,r = Metrikparameter

und für den Cosinus

$$\text{COSINUS}(i,j) = \frac{\sum_l g(i,j,l)X_{il}X_{jl})}{\left[\sum_l g(i,j,l)X_{il}^2 \sum_l g(i,j,l)X_{jl}^2\right]^{1/2}}$$

In der SPSS-X Prozedur CLUSTER ist das **paarweise Ausscheiden** als Option **nicht vorgesehen.** Eine einfache Abhilfe besteht darin, sich selbst ein kurzes Programm zur Berechnung von Ähnlichkeits- oder Unähnlichkeitsmatrizen zu schreiben. Als eine weitere praktikable Alternative kann das Programm QUICK CLUSTER angewendet werden. Das in

QUICK CLUSTER enthaltene Clusteranalyseverfahren ist im Anhang ausführlich dargestellt.

Die SPSS-X Programmeingabe für unser Beispiel wäre:

```
FILE HANDLE AFAMD / NAME - »AFAM.DAT«
GET FILE - AFAMD
MISSING VALUES AKERNF, AVERW, AINW, AGESIN ( - 9)
QUICK CLUSTER AKERNF AVERW AINW AGESIN
  /MISSING - PAIRWISE
  /CRITERIA - CLUSTERN)
```

Durch **die** MISSING VALUES Anweisung werden fehlende Werte für **die** Klassifikationsmerkmale AKERNF, AVERW, AINW und AGESIN definiert. Durch den Befehl QUICK CLUSTER wird die SPSS-X Prozedur aufgerufen. Unmittelbar daran anschließend werden die Klassifikationsmerkmale AKERNF, AVERW, AINW und AGESIN definiert. Das paarweise Ausscheiden wird durch die Anweisung MISSING = PAIRWISE vereinbart. In dem CRITERIA - Befehl wird die Anzahl der Cluster **vorgegeben.**

Bei dem QUICK-CLUSTER werden unmittelbar die Clustermittelwerte und - **Streuungen** ausgegeben. Das Verfahren eignet sich auch für große Datensätze, da im Unterschied zu den hierarchischen Verfahren keine Distanzmatrix für die Klassifikationsobjekte berechnet wird. Der Nachteil dieses Verfahrens besteht u.a. darin, daß nur die euklidische Metrik als Unähnlichkeitsmaß verwendet wird und deshalb implizit eine Gewichtung der Abstände in den einzelnen Klassifikationsmerkmalen durchgeführt wird (vgl. Abschnitt 2.3 und 2.4).

4.1.3 Mittelwertsubstitution

Dieses Verfahren wurde ebenfalls aus der Korrelationsanalyse übernommen. Die Grundidee besteht darin, die fehlenden Werte durch die Mittelwerte der Klassifikationsobjekte zu ersetzen. Das konkrete Vorgehen ist in der Tabelle 4.1-2 dargestellt.

Die Mittelwertsubstitution trifft bestimmte Annahmen über die Klassifikationsobjckte. Zunächst muß quantitatives Meßniveau vorliegen. Das fallweise oder paarweise Ausscheiden dagegen kann für Klassifikationmerkmale beliebigen Meßniveaus angewendet werden. Darüber hinaus ist die Anwendung der Mittelwertsubstitution nur sinnvoll, wenn die Klassifikationsmerkmale gleich »schwierig« sind und etwas »gemeinsames« messen, ihnen also latente Merkmale zugrundeliegen. Der Begriff der Schwierigkeit ist aus der psychologischen Testtheorie übernommen, inhaltlich bedeutet gleiche Schwierigkeit z.B., daß ein Test zur Messung der mathematischen Begabung aus gleich schwierigen Fragen besteht. Formal besagt die Annahme, daß die Erwartungswerte der Klassifikationsobjekte

identisch sind. Erwartungswerte stellen nun aber theoretische Größen dar. Würden allerdings keine fehlenden Werte vorliegen, wären die empirisch berechneten Mittelwerte brauchbare Schätzer für diese Erwartungswerte. Bei fehlenden Werten muß dieses aber nicht mehr der Fall sein, da fehlende Werte nicht zufällig auftreten müssen. Probleme entstehen vor allem dann, wenn gleiche Schwierigkeiten der Klassifikationsmerkmale nicht erwartet werden können. In diesem Fall wäre eine Zentrierung um die Erwartungswerte erforderlich, für die aber brauchbare Schätzer fehlen.

Tabelle 4.1-2:

Rechenschema für die Mittelwertsubstitution bei fehlenden Werten für das fiktive Beispiel der Tabelle 4.1-1.

Klassifikations-objekte	Klassifikationsmerkmale: AKERNF	AVERW	AINW	AGESIN	Mittelwerte
HH 1	1.00	0.00	0.00	3.00	1.00
HH 2	3.00	3.00	-9	0.00	2.00
HH 3	6.00	-9	0.00	0.00	2.00
HH 4	3.00	2.00	0.00	-9	1.67
HH 5	3.00	0.00	0.00	0.00	0.75
HH 1	1.00	0.00	0.00	0.00	
HH 2	3.00	3.00	2.00	0.00	
HH 3	6.00	2.00	0.00	0.00	
HH 4	3.00	2.00	0.00	1.67	
HH 5	3.00	0.00	0.00	0.00	

Die Annahme von zugrundeliegenden latenten Dimensionen ist in der Abbildung 4.1-1 dargestellt. In diesem Beispiel würde man einen fehlenden Wert in den Klassifikationsmerkmalen X1, X2 oder X3 durch den Mittelwert über die gültigen Ausprägungen der Klassifikationsmerkmale X1, X2 und X3 ersetzen.

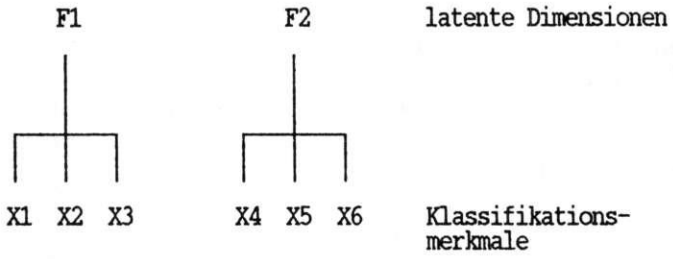
Technisch kann eine Mittelwertsubstitution für das Beispiel der Abbildung 4.1-1 in SPSS-X folgendermaßen durchgeführt werden:

```
FILE HANDLE XDAT/NAME = »X.DAT«
GET FILE - XDAT
MISSING VALUES X1 TO X6 ( - 9)
COMPUTE MX1«MEAN.1(X1,X2,X3)
```

```
COMPUTE MX2«MEAN.1(X4,X5,X6)
DO REPEAT X « X1 TO X3
IFMISS1NG(X)X«MX1
END REPEAT
DO REPEAT X - X4 TO X6
IF MISSING(X)X«MX2
END REPEAT
```

Abbildung 4.1-1:

Klassifikationsmerkmale als Indikatoren latenter Dimensionen



Durch die FILE HANDLE - Anweisung wird die SPSS·X Datei XDAT definiert, die extern unter dem Namen X.DAT abgespeichert ist. Diese Datei soll u.a. die Klassifikationsmerkmale X1, X2, ... ,X6 enthalten und wird durch die GET FILE - Anweisung »geladen«. Die Ausprägungen fehlender Werte werden durch die MISSING VALUES - Anweisung festgelegt. In dem Beispiel wurden fehlende Werte mit - 9 verkodet. Durch die Anweisung COMPUTE MX1 = MEAN.1 (X1,X2,X3) wird der Mittelwert in den Klassifikationsmerkmalen X1, X2 und X3 berechnet. Die Spezifikation MEAN.K führt allgemein dazu, daß der Mittelwert nur berechnet wird, wenn K und mehr gültige Ausprägungen gegeben sind. In den folgenden DO REPEAT - Schleifen werden fehlende Werte in den Klassifikationsmerkmalen durch die entsprechenden Mittelwerte ersetzt.

4.1.4 Vergleich der drei Verfahren

Jedes der drei dargestellten Verfahren besitzt bestimmte Vor - und Nachteile, die bei der Entscheidung für ein bestimmtes Vorgehen zu berücksichtigen sind (vgl. Tabelle 4.1-3).

Die Tabelle enthält nur das Zutreffen von Nachteilen der drei Verfahren. Treffen diese für ein Verfahren nicht zu, stellen sie Vorteile des entsprechenden Verfahrens dar. **Verzerrung der Repräsentativität der Klassifikationsobjekte** entsteht dann, wenn bei der fallweisen Elimination Klassifikationsobjekte ausgeschieden werden und die **fehlenden Werte nicht zufällig** verteilt sind. Analog dazu liegt eine **Verzerrung der Repräsentativität der Klassifikationsmerkmale** vor, wenn die Klassifikationsmerkmale **andere** als die theoretisch postulierten **latenten Dimensionen messen** und deshalb »falsche« Mittelwerte berechnet werden. Beim **paarweisen Ausscheiden** schließlich gehen in die Berechnung der Un - oder Ähnlichkeit zwischen zwei Klassifikationsobjekten nur jene Klassifikationsmerkmale ein, die in beiden Klassifikationsobjekten gültige Werte besitzen. Die **Klassifikationsmerkmale** können deshalb für jedes Paar von Klassifikationsobjekten **variieren.**

Tabelle 4.1-3:

Vor- und Nachteile der Behandlung fehlender Werte durch Mittelwertsubstitution, fallweise - oder paarweise Elimination

Nachteile:	Verfahren zur Behandlung fehlender Werte:		
	fallweise Elimination	paarweise Elimination	Mittelwert - substitution
Verzerrung d. Repräsentativität d.			
- Kl.objekte	ja	-	-
- Kl.merkmale	-	-	ja
- Ähn. - bzw. Unähnl.maße	-	ja	-
quantitatives Meßniveau erforderlich	-	-	ja
zusätzliche Modellannahmen	-	-	ja

Allerdings ist bis jetzt noch weitgehend ungeklärt, wie stark sich diese Fehler auf das Klassifikationsergebnis auswirken. Erste Simulationsstudien (Bacher 1987, Kaufman 1985) zeigen, daß das fallweise Ausscheiden zu einer geringeren Anzahl von Fehlklassifikationen führt als die Mittelwertsubstitution und die Mittelwertsubstitution wiederum zu geringeren Fehlzuordnungen als das paarweise Ausscheiden.

Mit aller Vorsicht, die bei der Verallgemeinerung von Simulationsexperimenten geboten ist, empfiehlt sich für die Forschungspraxis folgendes: Liegen nur wenige fehlende Werte vor, wird man die fallweise Elimination verwenden. Entsteht dadurch ein beträchtlicher Datenverlust, empfiehlt sich das Verfahren der Mittelwertsubstitution, sofern die Anwendungsvoraussetzungen erfüllt sind. Andernfalls wird man auf das paarweise Ausscheiden zurückgreifen.

Übungsaufgabe **14:**

a) Berechnen sie für die Tabelle 4.1-1 die Distanzmatrix (City-Blockmetrik) für die Methode des paarweisen Ausscheiden.

b) Sind in dem Rechenbeispiel der Tabelle 4.1-2 die Anwendungsvoraussetzungen für die Mittelwertsubstitution erfüllt? (Begründen Sie Ihre Antwort!)

c) Welchen Wert besitzen XI, X2 und X3 in der Abbildung 4.1-1, wenn alle Klassifikationsmerkmale einen fehlenden Wert aufweisen?

4.2 Transformation von Klassifikationsmerkmalen

In den bisherigen Ausführungen und Rechenbeispielen wurde immer angenommen, daß Vergleichbarkeit der Klassifikationsmerkmale vorliegt. In der Forschungspraxis wird aber diese Anwendungsvoraussetzung in zahlreichen Fällen nicht gegeben sein. In Abschnitt 2.3 wurden bereits Gründe für die Nichtvergleichbarkeit angeführt und Lösungsverfahren im Überblick dargestellt. Diese sollen nun detailliert behandelt werden.

4.2.1 Theoretische Gewichtung

Mit der theoretischen Gewichtung können unterschiedliche Probleme der Nichtvergleichbarkeit gelöst werden. Technisch kann sie auf die Klassifikationsmerkmale oder auf die Berechnung der Ähnlichkeit bzw. Unähnlichkeit angewendet werden. Beide Operationen lassen sich für die Minkowskimetrik ineinander überführen (10). Es ist deshalb vollkommen ausreichend, sich mit der theoretischen Gewichtung der Klassifikationsmerkmale zu beschäftigen. Sie kann angewendet werden bei:

- unterschiedlichen Maßeinheiten der Klassifikationsmcrkmale,
- Über - bzw. Unterrepräsentativität von Klassifikationsmerkmalcn,
- hierarchischen Klassifikationsmerkmalen (s. dazu Abschnitt 4.2.4) und für
- die bewußte Steuerung des Klassifikationsprozesses.

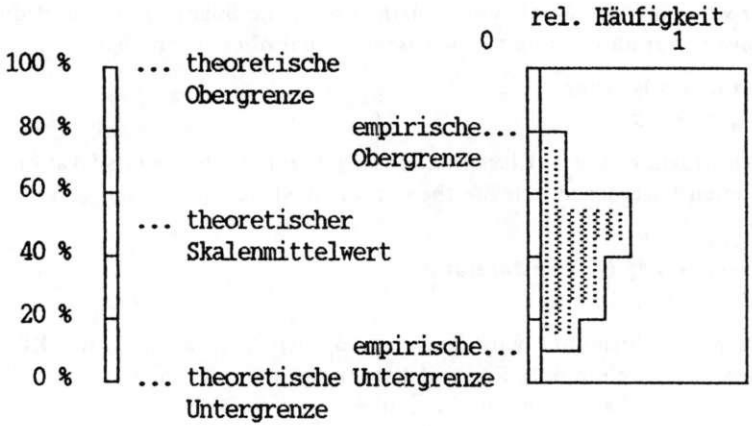
4.2.1.1 Theoretische Gewichtung bei unterschiedlichen Maßeinheiten

Vorausgesetzt werden zunächst quantitative Klassifikationsmerkmale X_1, X_2,... , für die theoretische Unter - (α_i) und Obergrenzen (β_i) sowie theoretische Skalenmittelwerte (μ_i) und -Streuungen (σ_i) bekannt sind. Bei diesen Größen handelt es sich um a priori bekannte Maßzahlen der verwendeten Skalen, in denen die Klassifikationsmerkmale gemessen werden. Sie können also ohne das Vorliegen empirischer Beobachtungen bestimmt werden. Zur Verdeutlichung des Unterschiedes zwischen diesen theoretischen Größen und ihren empirischen Pendants sei angenommen, daß Xj ein Anteilswert sei, der theoretisch zwischen 0 bzw. 0% und 1 bzw. 100% variieren kann. Die theoretische Untergrenze hat also auf jeden Fall den Wert 0 bzw. 0% und die theoretische Obergrenze den Wert 1 bzw. 100%, unabhängig von der konkreten empirischen Verteilung.

Der theoretische Skalenmittelwert beträgt 0.5 und die theoretische Skalenstreuung 0.25. Die empirisch beobachteten Werte können davon abweichen (vgl. Abbildung 4.2-1). In der Abbildung 4.2-1 beträgt z.B. die theoretische Obergrenze 80% und die empirische Untergrenze 10%.

Abbildung 4.2.-1:

Theoretische und empirische Skaleneinheiten



Unterschiedliche Maßeinheiten und folglich Nichtvergleichbarkeit liegen dann vor, wenn theoretische Unter - und Obergrenzen unterschiedliche Werte besitzen. Das bedeutet zugleich auch immer unterschiedliche

Skalenmittelwerte und Streuungen. Zur Beseitigung dieser Unvergleichbarkeit werden vier Verfahren angewendet:

Extremwertnormalisierung: Aus den Klassifikationsmerkmalen X_1, X_1, ... werden neue Klassifikationsmerkmale Z_1, Z_2, ... gebildet mit

$$Z_{il} = (X_{il} - \alpha_l)/(\beta_l - \alpha_l)$$

wobei

X_{il} = Ausprägung des Klassifikationsobjektes i in dem Klassifikationsmerkmal 1

Z_{il} = Ausprägung des Klassifikationsobjektes i in dem Klassifikationsmerkmal 1 nach der Transformation

Von der Ausprägung X_{il} wird zunächst die Untergrenze α_l subtrahiert und anschließend durch die Spannweite (Obergrenze minus Untergrenze) dividiert. Die so entstandenen neuen Klassifikationsmerkmale können zwischen 0 und 1 variieren.

Spannweitennormalisierung:

$$Z_{il} = X_{il}/\pi_l$$

mit

$\pi_l = \beta_l - \alpha_l$ = Spannweite

Bei diesem Vorgehen werden die Ausprägungen der Klassifikationsobjekte i nur durch die Spannweite (Obergrenze minus Untergrenze) dividiert. Bei der Anwendung der Minkowskimetrik führen die Spannweiten- und die Extremwertnormalisierung zu identischen Unähnlichkeitsmaßen **(11).**

Varianznormalisierung:

$$Z_{il} = X_{il}/\sigma_l$$

Die Ausprägungen der Klassifikationsobjekte i in den Klassifikationsmerkmalen 1 werden durch die theoretischen Skalenstreuungen (σ_l) dividiert.

Standardisierung (ZTransformation):

$$Z_{il} = (X_{il} - \mu_l)/\sigma_l$$

Bei der ZTransformation wird zunächst von den Ausprägungen der Klassifikationsobjekte i in dem Klassifikationsmerkmal 1 der theoretische Skalenmittelwert (μ_l) abgezogen und anschließend eine Division mit der theoretischen Skalenstreuung durchgeführt. Für die Minkowskimetrik ergeben sich für die Varianznormalisierung und die ZTransformation wiederum identische Unähnlichkeitswerte.

Die technische Durchführung dieser vier Operationen in SPSS X bereitet keine Probleme. Gegeben seien z.B. die Klassifikationsmerkmale X1

X2, X3 und X4, mit den theoretischen Untergrenzen ALPHAI, ALPHA2,, ALPHA3 und ALPHA4, den theoretischen Obergrenzen BETA 1, BETA2,, BETA3 und BETA4, den theoretischen Skalenmittelwerten MU1, MU2,, MU3 und MU4 und den theoretischen Skalenstandardabweichungen SIG-· MAI, S1GMA2, SIGMA3 und S1GMA4. Die entsprechenden SPSS X Ope-· rationen sind:

Extrem Wertnormalisierung:

```
COMPUTE ALPHAI = wert
COMPUTE ALPHA2 = wert
COMPUTE ALPHA3 = wert
COMPUTE ALPHA4 = wert
COMPUTE BETA1 = wert
COMPUTE BETA2 = wert
COMPUTE BETA3 = wert
COMPUTE BETA4 = wert
COMPUTE ZI - (X1 - ALPHA 1)/(BETA1 - ALPHAI)
COMPUTE Z2 - (X2 - ALPHA2 )/(BETA2 - ALPHA2)
COMPUTE Z3 - (X3 - ALPHA3)/(BETA3 - ALPHA3)
COMPUTE Z4 - (X4 - ALPHA4)/(BETA4 - ALPHA4)
```

Durch die Anweisung COMPUTE ALPHA1 = wert wird der theore-· tischen Untergrenze ALPHAI des Klassifikationsmerkmals X1 ein Zahlenwert zugewiesen, durch die Anweisung COMPUTE ALPHA2 = **wert** der theoretischen Untergrenze ALPHA2 des Klassifikationsmerkmals X2 ebenfalls ein Zahlenwert zugewiesen, usw... Analog wird in den COMPUTE - Anweisungen für die theoretischen Obergrenzen BETA1, BE-· TA2,... vorgegangen. In den darauf folgenden COMPUTE - Anweisungen werden die neuen Klassifikationsmerkmale Z1, Z2, Z3 und Z4 berechnet.

Spannweitennormalisierung:

```
COMPUTE P11 = wert d. obergrenze - wert d. untergrenze von X1
COMPUTE P12 = wert d. obergrenze - wen d. untergrenze von X2
COMPUTE P13 = wert d. obergrenze - wert d. untergrenze von X3
COMPUTE P14 = wert d. obergrenze - wert d. untergrenze von X4
COMPUTE Z1 = X1/P11
COMPUTE Z2 = X2/P12
COMPUTE Z3 = X3/P13
COMPUTE Z4 = X4/P14
```

In den ersten vier COMPUTE - Anweisungen werden den Variablen P11, P12, P13 und P14 die theoretischen Spannweiten der Klassifikation-· merkmale X1, X2, X3 und X4 zugewiesen. Die Transformation wird in den folgenden vier COMPUTE - Anweisungen durchgeführt.

Varianznormalisierung:

```
COMPUTE SIGMA1= wert
COMPUTE SIGMA2= wert
COMPUTE SIGMA3= wert
COMPUTE SIGMA4= wert
COMPUTE Z1 = X1/SIGMA1
COMPUTE Z2 = X2/SIGMA2
COMPUTE Z3 = X3/SIGMA3
COMPUTE Z4 = X4/SIGMA4
```

Standardisierung:

```
COMPUTE MU1= wert
COMPUTE MU2= wert
COMPUTE MU3= wert
COMPUTE MU4= wert
COMPUTE SIGMA1= wert
COMPUTE SIGMA2= wert
COMPUTE SIGMA3= wert
COMPUTE SIGMA4 = wert
COMPUTE Z1 = (X1 - MU1)/SIGMA1
COMPUTE Z2 = (X2 - MU2)/SIGMA2
COMPUTE Z3 = (X3 - MU3)/SIGMA3
COMPUTE Z4 = (X4 - MU4)/SIGMA4
```

Die **theoretische Gewichtung setzt voraus,** daß die **theoretischen Unter-** und **Obergrenzen** bzw. die **theoretischen Skalenmittelwerte** und **theoretischen Standardabweichungen** bekannt sind. Diese Anwendungsvoraussetzung ist beispielsweise in der sozialpsychologischen Einstellungsforschung erfüllt, wenn die Antworten auf unterschiedlichen Einstellungsskalen gemessen werden (Schlosser 1976: 56-58). Sie ist dagegen beipielsweise nicht erfüllt, wenn ein Teil der verwendeten Klassifikationsmerkmale in Prozentpunkten (Anteilswerten) und der andere Teil in absoluten Einheiten, wie z.B. das Bruttosozialprodukt in 1000,- DM je Einwohner gemessen wird. Für diese absoluten Einheiten fehlen in der Regel eindeutige theoretische Obergrenzen. Dieser Nachteil dürfte für die historische Sozialforschung in einem höheren Ausmaß zutreffen als für die Einstellungsforschung. Die theoretische Gewichtung dürfte daher für die Normierung der Skaleneinheiten in der historischen Sozialforschung weniger geeignet sein.

Sinnvoll und wertvoll ist die Anwendung der Spannweitennormalisierung bei ordinalen Dummy-Variablen, die bereits in Abschnitt 2.4.4 behandelt wurde. Bei ordinalen Klassifikationsmerkmalen sind aber i.d.R. empirische und theoretische Ober - und Untergrenzen identisch. Man könnte deshalb also auch von einer empirischen Spannweitennormalisierung sprechen.

4.2.1.2 Theoretische Gewichtung bei Über - bzw. Unterrepräsentativität von Klassifikationsmerkmalen

Über - bzw. Unterrepräsentativität liegt dann vor, wenn die Klassifikationsmerkmale gemeinsame, ihnen zugrundeliegende latente Dimensionen messen und jede dieser latenten Dimensionen durch eine unterschiedliche Anzahl von Indikatoren repräsentiert ist (s. Abschnitt 2.3). Die Gewichte werden in diesem Fall im Verhältnis zur Anzahl der Indikatoren vergeben: Wird z.B. die latente Dimension 1 durch 2 Indikatoren gemessen, die latente Dimension 2 durch 4 und die latente Dimension 3 durch 3 Indikatoren, dann erhalten die Indikatoren der latenten Dimension 1 das größte Gewicht mit 1/2, da diese Dimension im Vergleich zu den beiden anderen unterrepräsentiert ist, die der latenten Dimension 2 das kleinste Gewicht mit 1/4, da sie überrepräsentiert ist, und die der latenten Dimension 3 das Gewicht 1/3. Bei diesem Vorgehen wird nur die unterschiedliche Anzahl der Indikatoren, nicht aber deren Qualität berücksichtigt. Bei korrelierten Klassifikationsmerkmalen wird analog vorgegangen. Stark miteinander korrelierte Klassifikationsmerkmale erhalten kleinere Gewichte als schwach korrelierte.

4.2.1.3 Theoretische Gewichtung zur Steuerung des Klassifikationsprozesses

Oft ist eine explizite theoretische Steuerung des Klassifikationsprozesses erwünscht. Ein Beispiel dafür sind Klassifikationen von Verlaufskurven, bei denen bestimmten Zeitpunkten ein größeres Gewicht beigemessen werden soll (vgl. Blaschke/Liesegang 1977).

4.2.1.4 Theoretische oder empirische Gewichtung?

Bis auf die explizite Steuerung des Klassifikationsprozesses existiert zu jeder theoretischen Gewichtung ein empirisches Pendant, bei dem anstelle von theoretischen Gewichten empirische Gewichte verwendet werden (vgl. Abbildung 4.2-2).

Eine empirische Vorgehensweise hat allgemein zwei Nachteile, nämlich die Stichprobenabhängigkeit der Ergebnisse und die Abhängigkeit von Meßfehlern. Stichprobenabhängigkeit bedeutet, daß die empirischen Gewichte von der Verteilung der Untersuchungspopulation in den Klassifikationsmerkmalen abhängen. Dadurch entstehen Probleme beim Vergleich der Ergebnisse unterschiedlicher Untersuchungspopulationen, aber auch beim Vergleich einer Untersuchungspopulation zu mehreren Zeitpunkten. Betrachten wir z.B. folgende Situation: In eine Clusteranalyse wurden die Klassifikationsmerkmale Bruttosozialprodukt pro Kopf (BSP)

Abbildung 4.2-2:

Empirische und theoretische Gewichtung

Problem:	Lösungsverfahren: theoretisch	empirisch
unterschiedliche Maßeinheiten	Verwendung theoret. Ober- und Untergrenzen oder theoret. Skalenmittelwerte und -standardabweichungen	Verwendung emp. Ober- und Untergrenzen oder emp. Skalenmittelwerte u. -standardabweichungen
Unter bzw. Überrepräsentativität bzw. bei Korrelation der Kl.merkmale	Vergabe von theoret. Gewichten in Abhängigkeit von der Anzahl der Indikatoren. Die Qualität der Indikatoren geht in die Berechnung der Gewichte ein.	Berechnen von emp. Gewichten durch eine Faktorenanalyse. Die Qualität der ndikatoren geht in die Berechnung der Gewichte ein.

und der Industrialisierungsgrad (IG) einbezogen. Beide Klassifikationsmerkmale wurden zu zwei Zeitpunkten tl und t2 erhoben und für die Analyse empirisch standardisiert (siehe unten). Die Ergebnisse sind in der Tabelle 4.2-1 dargestellt. Es sind nun keine absoluten Aussagen, wie z.B. »in dem Cluster 3 hat sich das Bruttosozialprodukt in dem Zeitraum zwischen tl und t2 erhöht« mehr möglich, da z.B. in allen drei Clustern das Bruttosozialprodukt auf einen niederen Wert als dem des Clusters 1 zum Zeitpunkt tl gesunken **sein** kann. Nur mehr relative Aussagen sind möglich, wie z.B., daß sich **die** relativen Unterschiede der Cluster im Brutto**Sozialprodukt** verringert haben, während sie in der Industrialisierungsquote gestiegen sind.

Tabelle 4.2-1:

Fiktive Ergebnisse einer Clusteranalyse

Cluster	Klassifikationsmerkmale			
	BSP in t1	BSP in t2	IG in t1	IG in t2
Cluster 1	-.8	.0	.0	.3
Cluster 2	.0	.0	.0	.0
Cluster 3	.7	.0	.0	-.2

In die **empirische Gewichtung** gehen darüber hinaus die **Meßfehler** der Klassifikationsmerkmale ein. Das kann dazu führen, daß ein **Klassifikationsmerkmal künstlich »aufgebläht«** wird, obwohl die Unterschiede der Klassifikationsobjekte nur durch zufällige Meßfehler bedingt sind. Dieses Klassifikationsmerkmal mit nur zufälligen Unterschieden wird eine kleinere Standardabweichung besitzen als ein Klassifikationsmerkmal mit echten Unterschieden, das in derselben theoretischen Skaleneinheit gemessen wurde. Bei einer empirischen Standardisierung wird nun aber das Klassifikationsmerkmal mit nur zufälligen Unterschieden und einer kleinen Standardabweichung (durch die Division mit der Standardabweichung) stärker gewichtet als das Klassifikationsmerkmal mit echten Unterschieden (vgl. dazu auch Schlosser 1976: 56-89).

Auf der anderen Seite hat eine **theoretische Gewichtung** eben den **Nachteil,** daß die theoretischen Gewichte oft nicht definiert und/oder inhaltlich begründet sind. Insgesamt ergibt sich also eine Pattsituation, so daß allgemeine Handlungsregeln nicht gegeben werden können, außer der, daß beim Fehlen eindeutig inhaltlicher Kriterien mehrere Varianten durchgerechnet werden sollen.

Übungsaufgabe 15:

a) Beschreiben Sie die SPSS X Programme zur Varianznormalisierung und Z-Transformation
b) Bei der Varianznormalisierung und Z-Transformation ergeben sich für die Minkowskimetrik identische Unähnlichkeitsmaße. Überprüfen Sie diese Behauptung! Prüfen Sic ferner, **ob** sie auch für den Cosinus als Ähnlichkeitsmaß gilt?

4.2.2 Empirische Extremwert-, Spannweiten-, Varianznormalisierung und Standardisierung

Diese Operationen werden nach denselben Formeln wie ihre theoretischen Pendants durchgeführt. Der einzige Unterschied besteht in der Verwendung von empirischen anstelle von theoretischen Gewichten. Diese können z.B. in der SPSS X Prozedur FREQUENCIES (SPSS Inc. 1986:

314-328) bei der Verwendung der STATISTICS Option oder in der Prozedur CONDESCRIPTIVE (SPSS Inc. 1986: 328-336) berechnet werden. Für unser empirisches Beispiel der familialen Haushaltsdaten ergeben sich folgende empirische Maßzahlen:

Tabelle 4.2-2:

Empirische Maßzahlen der familialen Haushaltsstrukturen

Kl.merkmal	Mittelw.	Standardabw.	Obergr.	Untergr.
AKERNF	4.1	2.3	12	1
AVERW	1.9	2.3	9	0
AINW	.1	.3	2	0
AGESIN	.4	.9	5	0

Das Programm für die empirische Extremwertnormalisierung würde nun beispielweise folgendermaßen aussehen:

```
FILE HANDLE AFAMD / NAME = »AFAM.DAT«
GET FILE = AFAMD
COMPUTE ZKERNF=(AKERNF - 1.0)/11.0
COMPUTE ZVERW=(AVERW - 0.0)/9.0
COMPUTE ZINW=(AINW - 0.0)/2.0
COMPUTE ZGESIN=(ZGESIN -0.0)/5.0
CLUSTER ZKERNF ZVERW ZINW ZGESIN
  /....
```

In den COMPUTE - Anweisungen werden bei der Extremwertnormalisierung der Klassifikationsmerkmale neue Klassifikationsmerkmale ZKERNF, ZVERW, ZINW und ZGESIN erzeugt, die in die anschließende Clusteranalyse einbezogen werden können. Die Extremwertnormalisierung führt dazu, daß die Klassifikationsmerkmale mit einer geringen Spannweite, also in unserem Beispiel AINW und AGESIN stärker gewichtet werden. Dieser Effekt kann für die ersten fünf Haushalte unseres empirischen Beispiels verdeutlicht werden. Die Abbildung 4.2-3 enthält die Ausgangsdaten vor und nach einer Extremwertnormalisierung und die Ergebnisse des Complete-Linkage, wenn zur Messung der Unähnlichkeit die City-Blockmetrik verwendet wird. In diesem Beispiel führt die mit der Extremwertnormalisierung verbundene Aufwertung der Klassifikationsmerkmale mit kleiner Spannweite dazu, daß bei einer Clusteranalyse die Klassifikationsobjekte HH1 und HH2 stark von den anderen 3 Klassifikationsobjekten getrennt werden.

Diese Aufwertung tritt auch bei der ZTransformation für Klassifikationsmerkmale mit geringen Standardabweichungen ein. Insgesamt ergibt dabei eine Clusteranalyse (Complete-Linkage und City Blockmetrik) einen beträchtlichen Zuwachs bei einem Übergang von 5 zu 4 Clustern. Die entsprechenden Clustermittelwerte sind in der Tabelle 4.2-3 dargestellt.

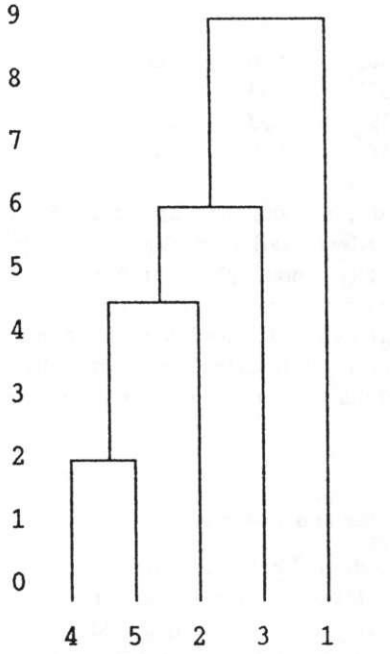
Abb. 4.2-3: Effekte der empirischen Extrerawertnoonalisierung

HH	AKERNF	AVERW	AINW	AGESIN
1	1	0	0	3
2	3	3	1	0
3	5	0	0	0
4	3	2	0	0
5	3	0	0	0

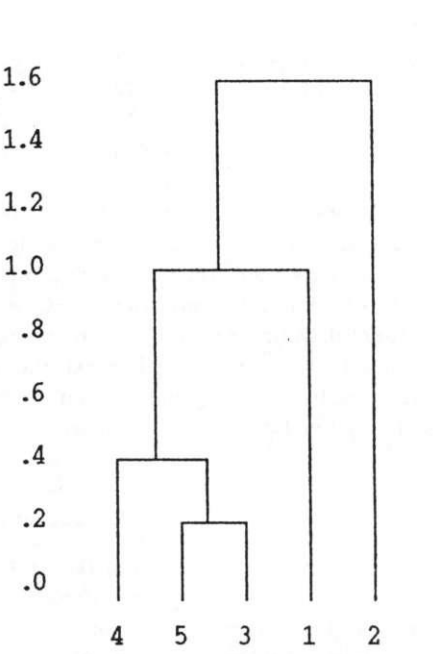
Ausgangsdaten vor der Extremwertnormalisierung

HH	ZKERNF	ZVERW	ZINW	ZGESBJ
1	.0	.0	.0	.6
2	.18	.33	.5	.0
3	.36	.0	.0	.0
4	.18	.22	.0	.0
5	.18	.0	.0	.0

Ausgangsdaten nach der Extremwertnormalisierung



Ergebnisse des Complete-Linkage vor der Extrem-wertnormalisierung



Ergebnisse des Complete-Linkage nach der Extrem-wertnormalisierung

Man sieht in dieser Tabelle, daß das 1. Cluster in dem Klassifikationsmerkmal ZKERNF, ZVERW und ZINW negative Werte besitzt, also Ausprägungen, die beträchtlich unter dem Durchschnitt der Untersuchungspopulation liegen. Dagegen besitzt es in dem standardisierten Klassifikationsmerkmal ZGESIN überdurchschnittlich hohe Ausprägungen. Das 2. Cluster dagegen ist durch einen hohen Wert in dem standardisierten Klassifikationsmerkmal ZINW gekennzeichnet, während die Ausprägung in dem Klassifikationsmerkmal ZVERW ungefähr dem Gesamtdurchschnitt der Untersuchungspopulation entspricht. Nochmals sei ausdrücklich betont, daß sich wegen der Standardisierung keine absoluten Unterschiede mehr festellen lassen. Aussagen der Art « im 4. Cluster treten mehr Verwandte (+ .038) als Mitglieder der Kernfamilie (-0.69) auf« sind unzulässig.

Tabelle 4.2-3:

Clustermittelwerte bei standardisierten Klassifikationsmerkmalen für familiale Haushaltsstrukturen

	ZKERNF	ZVERW	ZINW	ZGESIN
Cluster 1 (n = 7)	-.66	-.32	-.33	3.06
Cluster 2 (n - 13)	-.41	-.06	3.00	-.35
Cluster 3 (n = 70)	.74	-.37	-.33	-.05
Cluster 4 (n = 65)	-.69	.39	-.33	-.25
Cluster 5 (n = 4)	.28	.38	3.00	1.78

Vergleicht man diese 5-Clusterlösung mit der ursprünglichen 5-Clusterlösung ohne Standardisierung der Klassifikationsmerkmale, so ergibt sich nur eine geringe Übereinstimmung (symmetrisches Lambda = 0.22). Die extremwertnormalisierte 5-Clusterlösung zeigt eine noch geringere Übereinstimmung mit der ursprünglichen 5-Clusterlösung (symmetrisches Lambda = 0.07). Die extremwertnormalisierten und standardisierten 5-Clusterlösungen stimmen ebenfalls nur bescheiden überein (symmetrisches Lambda = 0.30).

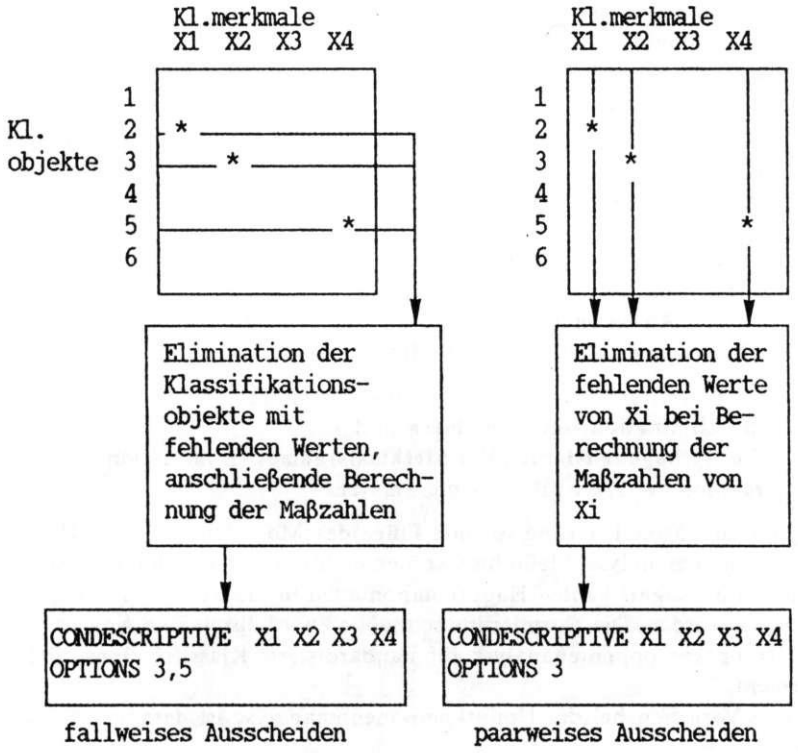
4.2.2.1 Behandlung fehlender Werte

Die Berechnung der empirischen Gewichte bei fehlenden Werten hängt von dem Verfahren ab, mit dem die fehlenden Werte behandelt werden. Bei der fallweisen Elimination müssen vorder Berechnung der Maßzahlen (Ober - , Untergrenze, Mittelwert und Standardabweichung) alle Klassifikationsobjekte mit fehlenden Werten eliminiert werden. Beim Verfahren des paarweisen Ausscheidens und bei der Mittelwertsubstitution werden

die fehlenden Werte nur spaltenweise eliminiert. Dieses Vorgehen ist in der Abbildung 4.2-4 dargestellt.

Abbildung 4.2-4:

Die Berechnung empirischer Gewichte bei fehlenden Werten



Ein Stern »*« symbolisiert in der Abbildung einen fehlenden Wert. Die Abbildung enthält ferner das entsprechende SPSS-X Programm CONDESCRIPTIVE (SPSS Inc. 1986: 328 - 335) zur Lösung dieser Aufgaben. Die Anweisung OPTIONS 3 bewirkt, daß die standardisierten Merkmalsausprägungen der Klassifikationsobjekte in neuen Klassifikationsmerkmalen zwischengespeichert werden. Diese neuen Klassifikationsmerkmale erhalten SPSS-X intern den Namen des ursprünglichen Klassifikationsmerkmals mit einem »Z« vorangestellt. Die zusätzliche Spezifikationsnummer 5 in der OPTIONS - Anweisung bewirkt fallweises Ausscheiden.

4.2.3 Die Hauptkomponentenanalyse

Die Hauptkomponentenanalyse oder Hauptachsentransformation wird in der Literatur oft als ein Verfahren der Faktorenanalyse dargestellt. Diese Einordnung ist aber nicht ganz zutreffend. Die **Hauptachsentransformation** ist ein **deskriptives Verfahren,** bei dem die Daten - oder Klassifikationsmatrix in einem **orthogonalen Raum** dargestellt wird. Die Achsen dieses orthogonalen Raumes werden als Hauptkomponenten bezeichnet. Der Faktorenanalyse dagegen liegt ein bestimmtes sozialpsychologisches oder psychologisches theoretisches Modell zugrunde, das eine kausale Beziehung zwischen den latenten Dimensionen (Faktoren) und den verwendeten Indikatoren unterstellt (12).

In SPSSX ist die Hauptkomponentenmethode ebenfalls in der Prozedur FACTOR enthalten, in der die in die Hauptkomponentenanalyse einbezogenen Variablen automatisch standardisiert werden. Tatsächlich kann die Hauptkomponentenanalyse aber auch für nichtstandardisierte oder nur mittelwertzentrierte Variablen angewendet werden (siehe dazu z.B. Sixtl 1982: 353-366).

Durch die **Anwendung** einer **Hauptkomponentenanalyse** werden folgende **Probleme** einer Klassifikationsaufgabe zu **lösen** versucht:

- die **Normierung** der Klassifikationsmerkmale,
- die **Elimination von Meßfehlern** und
- die **Orthogonalisierung des Merkmalsraumes** (Elimination der Korrelation der Klassifikationsmerkmale)

Die Orthogonalisierung ist eine Folge des Modellansatzes der Hauptkomponentenanalyse. Meßfehler werden dadurch zu eliminieren versucht, indem nur »signifikante« Hauptkomponenten in die weitere Analyse einbezogen werden. Die Normierung schließlich wird durch eine Anwendung der Hauptkomponentenanalyse auf standardisierte Klassifikationsobjekte erreicht.

Das Vorgehen bei der Hauptkomponentenanalyse ist demnach folgendes:
Die Klassifikationsmerkmale werden zunächst mit

$$Z_{il} = (X_{il} - m_l)/s_l$$

standardisiert, wobei

X_{il} = Ausprägung des Klassifikationsobjektes i in dem Klassifikationsmerkmal l
$m_l = (\Sigma X_{il})/n$ = Mittelwert des Klassifikationsmerkmals l
n = Anzahl der Klassifikationsobjekte
$s_l^2 = (\Sigma(X_{il} - m_l)^2)/n$ = Varianz des Klassifikationsmerkmals l
$s_l = \sqrt{s_l^2}$ = Standardabweichung des Klassifikationsmerkmals l

In einem zweiten Schritt werden nun die Hauptkomponenten gesucht. Sie stellen Linearkombinationen der Art

$$H_{ik} = a_{k1}Z_{i1} + a_{k2}Z_{i2} + \ldots + a_{km}Z_{im}$$

dar, mit

H_{ik} = Wert des Klassifikationsobjektes i in der k - ten Hauptkomponente

a_{kl} = Gewicht des Klassifikationsmerkmals l in der k - ten Hauptkomponente. (Die Gewichte werden in SPSSX als Faktorscores bezeichnet.)

Die Hauptkomponenten selbst werden schrittweise bestimmt. Dabei soll die erste Hauptkomponente die größte Varianz besitzen, die zweite Hauptkomponente die zweitgrößte, die dritte Hauptkomponente die drittgrößte, ...

Die Varianz einer Hauptkomponente ist gegeben durch:

$$\mathrm{Var}(H_k) = \Sigma H_{ik} = \Sigma(a_{k1}Z_{i1} + a_{k2}Z_{i2} + \ldots + a_{km}Z_{im})$$

$$\mathrm{Var}(H_k) = \Sigma H_{ik}^2 = \Sigma(a_{k1}Z_{i1} + a_{k2}Z_{i2} + \ldots + a_{km}Z_{im})^2$$

$$\mathrm{var}(H_1) > \mathrm{var}(H_2) > \ldots$$

mit

$$\mathrm{Var}(H_1) > \mathrm{Var}(H_2) > \ldots$$

Mathematisch läßt sich nun zeigen, daß die obige Maximierungsaufgabe - wie bei der Diskriminanzanalyse - zu einer Berechnung von Eigenwerten und Eigenvektoren führt. Die Eigenwerte entsprechen dabei den Varianzen der Hauptkomponenten. Für die Eigenwerte (Varianzen der Hauptkomponenten) gilt folgende Beziehung: Die Summe der Eigenwerte ergibt die Gesamtstreuung der Klassifikationsdatenmatrix. Diese ist für standardisierte Klassifikationsmerkmale identisch mit der Anzahl der Klassifikationsmerkmale. Die Varianz jeder Hauptkomponente kann deshalb zu dieser Gesamtvarianz in Beziehung gesetzt werden. Dadurch erhält man den Anteil erklärter Varianz. Ferner läßt sich zeigen, daß die Eigenwerte und Eigenvektoren aus der Korrelationsmatrix der Klassifikationsmerkmale berechnet werden können und die tatsächlichen Z-Werte also nicht benötigt werden.

Die Eigenvektoren enthalten die Gewichte a_{kl}. Der erste Eigenvektor enthält dabei die Gewichte der Klassifikationsmerkmale (k = 1, 2, .. . , m) in der ersten Hauptkomponente, der zweite Eigenvektor die Gewichte der Klassifikationsmerkmale k (k = 1, 2, . .. , m) der zweiten Hauptkomponente, usw. Es gilt folgende Beziehung:

1. Eigenwert = Varianz der 1. Hauptkomponenten = $a_{11}^2 + a_{21}^2 + \ldots a_{m1}^2$
2. Eigenwert = Varianz der 2. Hauptkomponenten = $a_{12}^2 + a_{22}^2 + \ldots a_{m2}^2$

Die Tabelle 4.2-4 enthält die Korrelationsmatrix der vier Klassifikationsmerkmale der familialen Haushaltsdaten.

Tabelle 4.2-4:

Korrelationsmatrix der Klassifikationsmerkmale der familialen Haushaltsdaten

	AKERNF	AVERW	AINW	AGESIN
AKERNF	1.00			
AVERW	-.20	1.00		
AINW	-.07	.03	1.00	
AGESIN	.03	-.03	.07	1.00

Die Klassifikationsmerkmale korrelieren insgesamt nur sehr schwach. Es ist deshalb zu erwarten, daß eine Hauptkomponentenanalyse kaum zu einer Reduktion des Merkmalsraumes beitragen wird. Das zeigt sich auch, wenn wir die Eigenwerte dieser Korrelationsmatrix betrachten:

Hauptkomponente	Eigenwert	%
1	1.22	30.7
2	1.06	26.7
3	.91	22.8
4	.79	19.8
	4.00	100.0

Es läßt sich **kein »signifikanter« Abfall** der Eigenwerte feststellen. Die Grundidee für die Bestimmung »signifikanter« Hauptkomponenten ist die Annahme, daß kleine Varianzen der Hauptkomponenten nur durch Meßfehler entstehen und deshalb bei einer Elimination dieser Hauptkomponenten mit kleinen Varianzen Meßfehler eliminiert werden können. In der Literatur haben sich zwei Verfahren zur Bestimmung der »signifikanten« Eigenwerte durchgesetzt: Das **Kaiserkriterium** und der Screetest. Beim Kaiserkriterium werden alle Hauptkomponenten mit einer Varianz kleiner 1.0 eliminiert, da diese weniger erklären als eine standardisierte Variable bzw. ein standardisiertes Klassifikationsmerkmal mit einer Varianz von 1.0. Beim Screetest werden die Eigenwerte graphisch dargestellt und - ähnlich wie bei der Bestimmung der Anzahl der Cluster - nach einem »signifikanten« Abfall für die Bestimmung der Anzahl der Hauptkomponenten gesucht. In der Abbildung 4.2- 6 würde dieses Vorgehen zu einer Auswahl von einer Hauptkomponente führen, hingegen läßt sich in

Abbildung 4.2-5:

Fiktive Beispiele für Screediagrarame

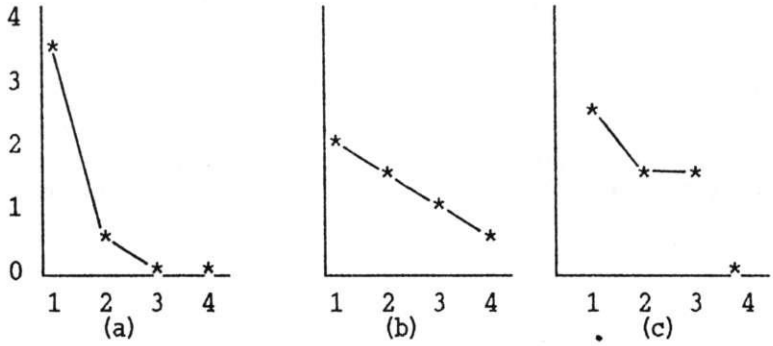
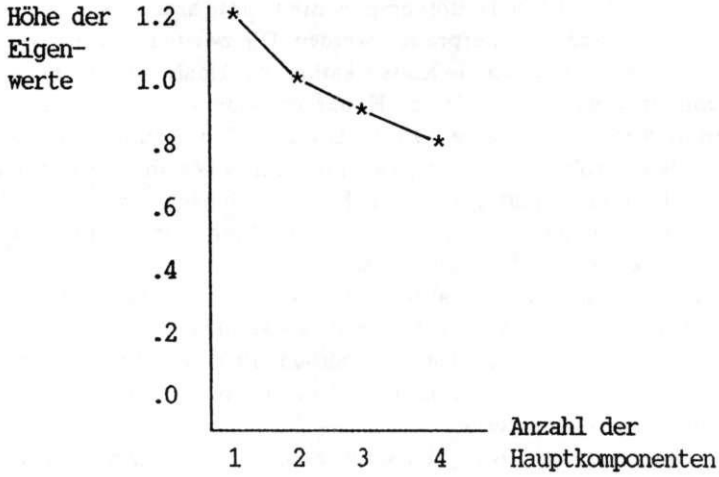


Abbildung 4.2-6:

Screediagramm für familiale Haushaltsdaten



der Abbildung 4.2-5b kein Abfall feststellen. Die Abbildung 4.2-5c schließlich enthält mehrere signifikante Abfälle.

In dem Beispiel der familialen Haushaltsdaten würden bei Anwendung des Kaiserkriteriums die dritte und vierte Hauptkomponente eliminiert werden. Dadurch bleibt aber insgesamt 40% der Gesamtvarianz ungeklärt. In dem Screediagramm dagegen läßt sich kein signifikanter Abfall feststellen (s. Abbildung 4.2-6).

Neben den Eigenwerten werden in demselben Rechenschritt die Eigen-**Vektoren** berechnet. Sie geben die Lage der Klassifikationsmerkmale in dem durch die Hauptkomponenten orthogonal aufgespannten Raum an. Die Eigenvektoren **für** unser Beispiel sind in der Tabelle 4.2-5 dargestellt.

Tabelle 4.2-5:

Eigenvektoren der familialen Haushaltsdaten

Klassifikations-	Hauptkomponenten:			
merkmale	1	2	3	4
AKERNF	-.77	.02	-.02	.64
AVERW	.73	-.13	.35	.57
AINW	.31	.69	-.63	.19
AGESIN	-.13	.76	.62	-.09

Die erste Hauptkomponente wird durch die beiden Klassifikationsmerkmale AKERNF und AVERW gebildet. Das negative Vorzeichen bedeutet eine Polarisierung. Auf einer Seite der Hauptkomponente liegt das Klassifikationsmerkmal AKERNF, auf der anderen das Klassifikationsmerkmal AVERW. Diese Hauptkomponente könnte also als »Kernfamilie versus Verwandtschaft« interpretiert werden. Die zweite und dritte Hauptkomponente werden durch die Klassifikationsmerkmale AINW und AGESIN gebildet, wobei bei der dritten Hauptkomponente wiederum eine Polarisierung beobachtet werden kann. Diese beiden Hauptkomponenten können als »Inwohner/Gesinde (Nichtgroßfamilie)« und »Inwohner versus Gesinde« interpretiert werden. Die Klassifikationsmerkmale AKERNF und AVERW bilden schließlich die vierte Hauptkomponente, die als »Großfamilie« interpretiert werden kann.

In vielen Fällen ist eine inhaltliche Interpretation der Hauptkomponenten nicht möglich. Man versucht deshalb die Hauptkomponenten zu drehen (rotieren), um eine eindeutig inhaltlich interpretierbare Lösung zu erhalten. Diese Rotationsverfahren sind in diesem Zusammenhang aber von keinem weiteren Interesse.

Die bisher dargestellten Ergebnisse wurden mit folgendem SPSS-X Programm berechnet:

```
FILE HANDLE AFAMD / NAME - »AFAM.DAT«
GET FILE - AFAMD
FACTOR VARIABLES - AKERNF AVERW AINW AGESIN
  /PRINT - CORRELATION INITIAL EXTRACTION
  /CRITERIA - FACTORS(4)
  /ROTATION - NOROTATE
  /EXTRACTION « PC
```

Durch die Anweisung FACTOR wird die SPSS-X Prozedur FACTOR aufgerufen. Die Variablen, die in die Analyse einbezogen werden sollen, werden in der VARIABLES - Anweisung definiert. Die PRINT - Anweisung bewirkt, daß die Korrelationsmatrix der Klassifikationsmerkmale (CORRELATION), alle Eigenwerte und Eigenvektoren (INITIAL), die berechnet werden können, und die Eigenwerte und Eigen Vektoren (EXTRACTION), die in die weitere Analyse einbezogen werden, ausgegeben werden (13). Die Anweisung CRITERIA - FACTORS(4) legt die Anzahl der Hauptkomponenten (-4) für die weitere Analyse, wie z.B. einer Rotation, fest. Die Voreinstellung ist die Bestimmung nach dem Kaiserkriterium. Soll diese Voreinstellung aufgehoben werden, muß eine CRITERIA - Anweisung vor dem EXTRACTION - Befehl geschrieben werden. Durch ROTATION = NOROTATE wird festgelegt, daß keine Rotation durchgeführt werden soll (14). Die Voreinstellung wäre eine VARIMAX-Rotation. Die Methode zur Bestimmung der »Faktoren« wird in dem EXTRACTION - Befehl definiert. PC bedeutet die Anwendung einer Hauptkomponentenmethode (Principal Components).

In der Ausgabe werden noch zusätzlich die sogenannten Kommunalitäten ausgedruckt. Diese erhält man dadurch, indem die Elemente der Matrix der Eigenvektoren quadriert und zeilenweise aufaddiert werden. In unserem Beispiel betragen die Kommunalitäten 1.0. Sie geben an, wieviel % der Varianz durch die Hauptkomponenten erklärt werden. (Berechnet man dagegen für die quadrierten Werte die Spaltensummen, erhält man die Eigenwerte.)

Durch dieses Vorgehen erhält man alle für die Berechnung der Gewichte zur Bestimmung der Werte der Klassifikationsobjekte in Hauptkomponenten erforderlichen Informationen. Die Werte der Klassifikationsobjekte in den Hauptkomponenten können nach vier Methoden berechnet werden (vgl. Kauf man 1985).

1. Berechnung aller ungewichteten Hauptkomponenten.

Bei dieser Methode werden die Gewichte so berechnet, daß die Varianzen der Hauptkomponenten gleich 1.0 gesetzt werden. Man führt also eine Standardisierung der Hauptkomponenten durch. Die Gewichte ergeben sich dadurch, daß die Elemente des 1 - ten Eigenvektors durch die Standardabweichung (Wurzel des Eigenwertes) der 1 - ten Hauptkompo-

nente dividiert werden usw.. Für unser Beispiel führt dieses Vorgehen dazu, daß die erste Spalte der Tabelle 4.2-6 durch $\sqrt{1.22}$ dividiert wird, die zweite Spalte mit $\sqrt{1.06}$ usw. Als Ergebnis erhält man folgende Matrix der Gewichte (vgl. Tabelle 4.2-6).

Nach der Formel

$$H_{ik} = a_{1k}Z_{i1} + a_{2k}Z_{i2} + ... + a_{mk}Z_{im}$$

lassen sich nun die Werte der Klassifikationsobjekte i in den Hauptkomponenten k berechnen. Für den ersten Haushalt mit den standardisierten Merkmalsausprägungen ZKERNF - -1.35, ZVERW - -.83, ZINW = -.25 und ZGESIN - 2.89 ergeben sich folgende Werte:

$$H_{11} = -.70*(-1.35) + .66*(-.83) + .28*(-.25) - .12*(2.89) = -0.02$$
$$H_{12} = .02*(-1.35) - .13*(-.83) + .68*(-.25) + .75*(2.89) = 2.08$$
$$H_{13} = -.02*(-1.35) + .37*(-.83) + -.66*(-.25) + .65*(2.89) = 1.76$$
$$H_{14} = .72*(-1.35) + .64*(-.83) + .20*(-.25) + -.10*(2.89) = -1.84$$

Tabelle 4.2-6:

Matrix der Gewichte a_{kl} für alle ungewichteten Hauptkomponenten

Klassifikations-	Hauptkomponenten:			
merkmale	1	2	3	4
AKERNF	-.70	.02	-.02	.72
AVERW	.66	-.13	.37	.64
AINW	.28	.68	-.66	.20
AGESIN	-.12	.75	.65	-.10

2. Berechnung aller gewichteten Hauptkomponenten.

Bei diesem Vorgehen werden die Elemente der Eigen**Vektoren** unmittelbar als Gewichte verwendet. Im Unterschied zu der ungewichteten Vorgehensweise kommt deshalb den Hauptkomponenten mit einer größeren Varianz (Eigenwert) ein größeres Gewicht zu.

3. Berechnung der signifikanten ungewichteten Hauptkomonenten.

Bei diesem Verfahren werden nur die Werte der Klassifikationsobjekte in den »signifikanten« Hauptkomponenten berechnet, die z.B. nach dem Kaiserkriterium bestimmt werden. Für diese »signifikanten« Hauptkomponenten werden die Gewichte wie in Punkt 1 berechnet.

4. Berechnung der signifikanten gewichteten Hauptkomponenten.

Es werden nur die Werte der Klassifikationsobjekte in den »signifikanten« Hauptkomponenten wie in Punkt 2 berechnet.

Simulationsexperimente (Kaufman 1985) haben gezeigt, daß die Berechnung von ungewichteten Hauptkomponenten extrem gegenüber Meßfehlern anfällig ist. Darüber hinaus erhöht sich der Anteil von Fehlklassifikationen, wenn nur signifikante Hauptkomponenten in die Analyse einbezogen werden. Das bedeutet aber, daß durch die Auswahl »signifikanter« Hauptkomponenten »Meßfehler« kaum reduziert werden. Es empfiehlt sich also dies als Methode 2 beschriebene Verfahren (Berechnung aller gewichteten Hauptkomponenten).

Technisch läßt sich die Berechnung der Hauptkomponenten wie folgt realisieren:

```
FILE HANDLE AFAMD / NAME = »AFAMD.DAT«
GET FILE = AFAMD
CONDESCRIPTIVE AKERNF AVERW AINW AGESIN
OPTIONS 3
COMPUTE H1 = ZAKERNF*(-.70) + ZAVERW*(.66) + ZAGESIN*(.28) +
             ZAINW*(-.12)
COMPUTE H2 = ZAKERNF*(-.02) + ZAVERW*(.13) + ZAGESIN*(.68) +
             ZAINW*(.75)
COMPUTE H3 = ZAKERNF*(-.02) + ZAVERW*(.37) + ZAGESIN*(-.66) +
             ZAINW*(.65)
COMPUTE H4 = ZAKERNF*(.72) + ZAVERW*(.64) + ZAGESIN*(.20) +
             ZAINW*(-.10)
CLUSTER H1 H2 H3 H4
  /....
```

In diesem Beispiel werden alle ungewichteten Hauptkomponenten berechnet. Die Klassifikationsmerkmale werden in der Prozedur CONDESCRIPTIVE durch die OPTION Anweisung standardisiert und in den Variablen ZAKERNF, ZAVERW, ZAINW und ZAGESIN zwischengespeichert. Dabei wird der ursprünglichen Variablenbezeichnung ein Z vorangestellt. Zu beachten ist, daß der letzte Buchstabe einer Variablen abgeschnitten wird, wenn diese bereits die maximale Länge von 8 Zeichen besitzt. In den darauf folgenden COMPUTE - Anweisungen werden die Werte der Klassifikationsobjekte in den Hauptkomponenten berechnet. Diese Werte werden als neue Klassifikationsmerkmale in die Clusteranalyse einbezogen.

4.2.3.1 Behandlung fehlender Werte in der Hauptkomponentenanalyse

Fehlende Werte müssen bei der Hauptkomponentenanalyse zweimal bearbeitet werden:

Erstens bei der Berechnung der Korrelationsmatrix und zweitens bei der Berechnung der Werte der Klassifikationsobjekte in den Hauptkomponenten. Bei der Berechnung der Korrelationsmatrix kann die Methode der fallweisen-, oder der paarweisen Elimination sowie die Mittelwertsub-

stitution angewendet werden. Die dafür notwendigen Operationen beziehen sich - im Unterschied zu Kapitel 4.1 - auf die Spalten der Klassifikationsdatenmatrix. Bei der Mittelwertsubstitution wird also anstelle des Mittelwertes eines Klassifikationsobjektes der Mittelwert des Klassifikationsmerkmales für fehlende Werte eingesetzt.

Im zweiten Schritt (bei der Berechnung der Werte der Klassifikationsobjekte in den Hauptkomponenten) wird dasselbe Verfahren wie bei der Berechnung der Korrelationsmatrix angewendet. Das Vorgehen ist demnach folgendes:

Fallweises Ausscheiden:

```
CONDESCRIPTIVE AKERNF AVERW AINW AGESIN
OPTIONS 3 5
COMPUTE H1 = .....
COMPUTE H2 = .....
....
CLUSTER H1 H2 H3 H4
  /MISSING = LISTWISE
```

Die OPTIONS - Anweisung, die an die CONDESCRIPTIVE - Prozedur anschließt, bewirkt **fallweises Ausscheiden.** Jedes Klassifikationsobjekt,

```
CONDESCRIPTIVE AKERNF AVERW AINW AGESIN
OPTIONS 3
COMPUTE AA=NVALID(ZAKERNF, ZAVERW, ZAINW, ZAGESIN)
RECODE ZAKERNF ZAVERW ZAINW ZAGESIN (MISSING = 0)
COMPUTE H1=....
COMPUTE H2=....
COMPUTE H3=....
COMPUTE H4=....
COMPUTE H1 = (4.0/AA)*H1
COMPUTE H2 = (4.0/AA)*H2
COMPUTE H3 = (4.0/AA)*H3
COMPUTE H4 = (4.0/AA)*H4
CLUSTER H1 H2 H3 H4
  /...
```

```
CONDESCRIPTIVE AKERNF AVERW AINW AGESIN
OPTIONS 3
COMPUTE AA=NVALID(ZAKERNF, ZAVERW, ZAINW, ZAGESIN)
RECODE ZAKERNF ZAVERW ZAINW ZAGESIN (MISSING = 0)
COMPUTE H1=....
COMPUTE H2=....
COMPUTE H3=....
COMPUTE H4=....
COMPUTE H1 = (4.0/AA)*H1
COMPUTE H2 = (4.0/AA)*H2
COMPUTE H3 = (4.0/AA)*H3
COMPUTE H4 = (4.0/AA)*H4
CLUSTER H1 H2 H3 H4
  /...
```

In der Prozedur CONDESCRIPTIVE wird der standardisierte Wert für jedes Klassifikationsmerkmal getrennt berechnet. Die fehlenden Werte der standardisierten Klassifikationsmerkmale treten also an derselben Stelle wie bei den unstandardisierten auf. (Sie erhalten intern den Wert SYSMIS.) In der Anweisung COMPUTE AA - NVALID(ZAKERNF, ZA-· VERW, ZAINW, ZAGESIN) wird für jedes Klassifikationsobjekt die Anzahl gültiger Werte gezählt. In der anschließenden RECODE - Anweisung werden die fehlenden Werte auf 0 gesetzt. Dadurch beeinflussen sie die in den folgenden vier COMPUTE - Anweisungen berechneten Werte in den Hauptkomponenten nicht. (Diese erhalten wegen der Umkodierung keinen fehlenden Wert, wenn in den Z-Variablen eine oder mehrere Ausprägungen fehlen.) In den letzten vier COMPUTE Anweisungen werden die Werte der Hauptkomponenten mit 4.0/AA (4.0 = Anzahl der Klassifikationsmerkmale, AA = Anzahl gültiger Ausprägungen) reskaliert.

Anmerkung: Besitzt ein Klassifikationsobjekt in allen Klassifikationsmerkmalen einen fehlenden Wert, dann ist AA = 0 und die Division nicht möglich. In diesem Fall können vor der Reskalierung durch die Anweisung SELECT IF (AA GT 0) diese Klassifikationsobjekte für die weitere Analyse eliminiert werden.

Mittelwertsubstitution:

```
CONDESCRIPTIVE AKERNF AVERW AINW AGESIN
OPTIONS 3
COMPUTE ZM - MEAN.KZAKERNF, ZAVERW, ZAGESIN, ZAINW)
IF MISSING(ZAKERNF) ZAKERNF = ZM
IF MISSING(ZAVERW) ZAVERW = ZM
IF MISSING(ZAGESIN) ZAGESIN = ZM
IF MISSING(ZAINW) ZAINW = ZM
COMPUTE Hl - ...
COMPUTE H2 - ...
COMPUTE H3 «= ...
COMPUTE H4 - ...
CLUSTER Hl H2 H3 H4
  /MISSING - LISTWISE
```

Durch die Anweisung ZM - MEAN.1(ZAKERNF, ZAVERW, ZAINW, ZAGESIN) wird für jedes Klassifikationsobjekt der Mittelwert in den standardisierten Klassifikationsmerkmalen ZAKERNF, ZAVERW, ZAINW und ZAGESIN berechnet und der Variablen ZM zugewiesn. Die Anweisung IF MISSING(ZAKERNF) ZAKERNF=ZM bewirkt, daß in ZAKERNF der Mittelwert ZM eingesetzt wird, wenn ein Klassifikationsobjekt in ZAKERNF einen fehlenden Wert besitzt. Diese Operation wird für alle Klassifikationsmerkmale wiederholt. Anschließend werden die Hauptkomponenten berechnet und eine Clusteranalyse durchgeführt.

Anmerkung: Hat ein Klassifikationsobjekt in allen Klassifikationsmerkmalen fehlende Werte, hat ZM ebenfalls einen fehlenden Wert. In diesem Fall ist keine Mittelwertsubstitution möglich und wird in SPSSX auch nicht durchgeführt. Als Konsequenz können auch die Hauptkomponenten fehlende Werte aufweisen. Deshalb empfiehlt sich in der Prozedur **Cluster** die Anwendung der Anweisung MISSING = LISTWISE.

Übungsaufgabe 16:

a) Welches Meßniveau müssen die Klassifikationsmerkmale besitzen, damit eine Hauptkomponentenanalyse durchgeführt werden kann? Begründen Sie Ihre Antwort!

b) In einer Hauptkomponentenanalyse wurden folgende Eigenvektoren berechnet. Berechnen Sie die dazugehörenden Eigenwerte.

	1	2
BSP je Einw.	.45	.33
Industrialisierungsgrad	.88	.20
Schuldendienst	.77	-.23
Alphabetisierungsquote	.13	.69
Geburtenrate	.23	.60
Anteil Ärzte je Einw.	.12	.72

c) Schreiben Sie ein SPSS X Programm, das für die familialen Haushaltsdaten alle gewichteten Hauptkomponenten berechnet.

4.2.4 Die Berechnung geeigneter Unähnlichkeitsmaße

4.2.4.1 Die Mahalanobisdistanz

Die Mahalanobisdistanz wird bei quantitativen Klassifikationsmerkmalen zur Elimination der Korrelation der Klassifikationsmerkmale verwendet. Die allgemeine Formel zur Berechnung der Mahalanobisdistanz MD(iJ) zwischen zwei Klassifikationsobjekten i und j ist:

$$MD(i,j) = \Sigma \Sigma (X_{il} - X_{jl}) s^{lk} (X_{ik} - X_{jk})$$

mit

X_{il} = Ausprägung des Klassifikationsobjektes i in dem Klassifikationsmerkmal 1

s^{lk} = Element l,k der Inversen der Kovarianzmatrix der Klassifikationsmerkmale.

Die Formel sieht zunächst etwas unübersichtlich aus. Für den Fall, daß unkorrelierte Klassifikationsmerkmale vorliegen, daß also $s^{lk} = 0$ für alle $1 \neq k$ gilt, vereinfacht sich die Formel zu

$$MD(i,j) = \Sigma s^{ll} (X_{il} - X_{jl})^2$$

Sie entspricht also der quadrierten euklidischen Distanz für varianznormalisierte oder standardisierte Klassifikationsmerkmale.

Bei **korrelierten Klassifikationsmerk malen** ist die Mahalanobisdistanz identisch mit der **quadrierten euklidischen Distanz** aller ungewichteten Hauptkomponenten, wenn die Hauptkomponenten für die Kovarianz· matrix (anstelle der Korrelationsmatrix) berechnet werden. Liegen also standardisierte Klassifikationsmerkmale vor, kann in SPSSX die Mahalanobisdistanz über den Umweg einer Hauptkomponentenanalyse berechnet werden. In die Clusteranalyse werden dabei alle ungewichteten Hauptkomponenten einbezogen und als Unähnlichkeitsmaß die quadrierte euklidische Distanz (MEASURE = SEUCLID) verwendet. Für unstandardisierte Klassifikationsmerkmale kann in SPSSX die Mahalanobisdistanz nicht berechnet werden, da in der Prozedur FACTOR keine Kovarianzmatrix analysiert werden kann und die Mahalanobisdistanz in der Prozedur PROXIMITY nicht als Unähnlichkeitsmaß vorgesehen ist.

4.2.4.2 Der Informationsradius von Jardine und Sibson

Der Informationsradius von Jardine und Sibson (1971: 10-20) kann in SPSS·X ebenfalls nicht berechnet werden. Ausgangspunkt des Informationsradius bilden Aggregate, die eine Verteilung in einer oder mehreren nominalen Klassifikationsmerkmalen besitzen. Diese Verteilungen lassen sich durch eine Aggregation von nominalen Dummies erzeugen. In Abschnitt 2.3 wurde gezeigt, daß die City-Blockmetrik als Unähnlichkeitsmaß für diese Art von Aggregaten verwendet werden kann. Der Informationsradius stellt eine Alternative zur City-Blockmetrik dar.

Für ein nominales Merkmal mit den Ausprägungen $A_1, A_2, .., A_m$ besitzen die Klassifikationsobjekte 1 und 2 die Verteilungen $P_1(A1), P_1(A2), ..., P_1(Am)$ und $P_2(A1), P_2(A2), ..., P_2(Am)$. Der Informationsradius I(1,2) ist definiert als:

$$I(1,2) = [\sum_i \sum_k P_i(Ak)\log_2(2P_i(Ak)/\sum_j P_j(Ak))]/2.0$$

Die Formel läßt sich umformen in:

$$I(1,2) = [\Sigma P_1(Ak)\log_2 \frac{P_1(Ak)}{(P_1(Ak) + P_2(Ak))/2} + \Sigma P_2(Ak)\log_2 \frac{P_2(Ak)}{(P_1(Ak) + P_2(Ak))/2}]/2.0$$

$$= [I(1/1+2) + I(2/1+2)]/2.0$$

I(1/1 + 2) stellt den Informationsgewinn dar, den die Verteilung des Klas·sifikationsobjektes 1 gegenüber der gemeinsamen Verteilung der Klassifikationsobjekte 1 und 2 enthält. I(2/1 + 2) ist analog definiert. Allgemein ist I(1/1 + 2) ungleich I(2/1-1-2). Beim Informationsradius wird deshalb das

arithmetische Mittel berechnet, um ein symmetrisches Unähnlichkeitsmaß zu erhalten. Der Informationsradius kann zwischen 0.0 (kein Informationsgewinn = Gleichheit der Verteilung) und 1.0 (maximaler Informationsgewinn) variieren.

Liegen mehrere Klassifikationsmerkmale vor, kann für jedes Klassifikationsmerkmal 1 der Informationsradius berechnet werden. Die Summe ergibt den Gesamtinformationsradius.

Die Berechnung soll anhand eines Beispiels demonstriert werden. Gegeben sind die beiden Verteilungen $P_1(A1) = .20$, $P_1(A2) = .40$, $P_1(A3) = .30$ und $P_1(A4) = .10$ sowie $P_2(A1) = .30$, $P_2(A2) = .40$, $P_2(A3) = .10$ und $P_2(A4) = .20$.

Für den Informationsradius I(1,2) erhält man:

$$\begin{aligned} I(1,2) = & -(.20\log_2.40/.50 + .30\log_2.60/.50 + \\ & .40\log_2.80/.80 + .40\log_2.80/.80 + \\ & .30\log_2.60/.40 + .10\log_2.20/.40 + \\ & .10\log_2.20/.30 + .20\log_2.40/.30)/2.0 \\ = & \ 0.057 \end{aligned}$$

Zum Vergleich beträgt die City-Blockmetrik

$$d(1,2) = Abs(.20 - .30) + Abs(.40 - .40) + Abs(.30 - .10) + Abs(.10 - .20) = .40$$

Der Informationsradius I(1,2) eignet sich insbesondere bei hierarchischen Merkmalen, dabei werden Gewichte für die einzelnen Anteilswerte eingeführt (vgl. Fox 1982). Konkrete Anhaltspunkte für die Bestimmung der Gewichte geben Jardine und Sibson und auch Fox nicht.

Der Informationsradius läßt sich - wie die Mahalanobisdistanz - in SPSS X nicht berechnen. Eine mit SPSS-X praktikable Lösung für hierarchische Merkmale sieht folgendermaßen aus:

Die Abbildung 4.2-7 enthält ein Beispiel für hierarchische Merkmale. Das Klassifikationsmerkmal B mit den Ausprägungen B1, B2 und B3 tritt nur auf, wenn die Ausprägung A1 des Klassifikationsmerkmals A gegeben ist; das Klassifikationsmerkmal C nur, wenn die Ausprägung B3 vorliegt. Der erste Schritt besteht darin, daß die Klassifikationsmerkmale in Dummies aufgelöst werden (vgl. Abbildung 4.2-7). Ein Stern bedeutet einen fehlenden Wert.

Für diese fehlenden Werte müssen nun Schätzwerte eingesetzt werden, um eine fallweise Elimination bei der Anwendung der Prozedur CLUSTER zu vermeiden. Man kann diese durch die Annahme einer Gleichverteilung auf den nichtdefinierten Ausprägungen gewinnen. Die fehlenden Werte für B werden folglich durch 1/3 ersetzt, die von C durch 1/2.

Das entsprechende SPSSX Programm besitzt folgende Struktur:

```
DO REPEAT DA=DA1 TO DA2/WERT = 1 TO 2
COMPUTE DA=0
IF (A = WERT) DA=1
END REPEAT
DO REPEAT DB=DB1 TO DB3/WERT = 1 TO 3
COMPUTE DB = 0
IF (B = WERT) DB = 1
IF MISSING(B) DB = 1./3.
END REPEAT
DO REPEAT DC=DC1 TO DC3/WERT = 1 TO 2
COMPUTE DC=0
IF (C = WERT) DC=1
IF MISSING(C) DC=1./2
END REPEAT
```

Abbildung 4.2-7:

Hierarchische Merkmale und ihre Auflösung in Dummies

A	B	C	a1	a2	b1	b2	b3	c1	c2
A1	B1		1	0	1	0	0	*	*
	B2		1	0	0	1	0	*	*
	B3	C1	1	0	0	0	1	1	0
		C2	1	0	0	0	1	0	1
A2			0	1	*	*	*	*	*

In der ersten DO REPE AT - Schleife werden die Dummy-Variablen von A gebildet, in der zweiten Schleife die Dummies von B. Die Anweisung IF MISSING(B) DB = 1./3. bewirkt, daß bei fehlender Information die Dummies DB1, DB2 und DB3 den Wert 1/3 erhalten. Für das Klassifikationsmerkmal C wird analog verfahren.

Anzumerken ist, daß bei diesem Vorgehen die Klassifikationsmerkmale gleich gewichtet werden. In bestimmten Anwendungssituationen soll aber die Hierarchie der Merkmale in die Klassifikation eingehen. In diesem Fall können die Dummies gewichtet werden: Die Dummies der übergeordneten Klassifikationsmerkmale, also von Klassifikationsmerkmalen, von

denen das Auftreten anderer Klassifikationsmerkmale abhängt, erhalten dabei ein höheres Gewicht. In unserem Beispiel könnten die Dummies von A mit 3.0, die von B mit 2.0 und die von C mit 1.0 gewichtet werden.

4.2.4.3 Das Distanzmaß von Gower

Das Unähnlichkeitsmaß von Gower(1970) wurde zur Messung der Unähnlichkeit von Klassifikationsobjekten bei Klassifikationsmerkmalen mit unterschiedlichem Meßniveau entwickelt. Betrachten wir beispielsweise folgende Ausgangskonstellation: Gegeben sind die Klassifikationsmerkmale:

X = quantitatives Klassifikationsmerkmal mit der Spannweite RX
Y = quantitatives Klassifikationsmerkmal mit der Spannweite RY
Q - ordinales Klassifikationsmerkmal mit den Ausprägungen Q1, Q2, Q3 und Q4
P = ordinales Klassifikationsmerkmal mit den Ausprägungen P1, P2 und P3
A = nominales Klassifikationsmerkmal mit den Ausprägungen A1, A2 und A3
B = nominales Klassifikationsmerkmal mit den Ausprägungen B1, B2 und B3.

Die allgemeine Idee des Unähnlichkeitsmaßes von Gower besteht nun darin, die Klassifikationsmerkmale so zu gewichten, daß jedem Klassifikationsmerkmal in der Messung der Unähnlichkeit dasselbe Gewicht zukommt. Wir wissen, daß bei der Verwendung der City-Blockmetrik der maximale Wert eines quantitativen Klassifikationsmerkmals gleich der Spannweite ist, bei ordinalen gleich der Anzahl der Ausprägungen minus 1 und bei nominalen gleich 2 ist. Damit sind aber unmittelbar die Gewichte bekannt. Quantitative Klassifikationsmerkmale werden mit dem inversen Wert der Spannweite, ordinale mit dem inversen Wert der Anzahl der Ausprägungen minus 1 und nominale mit dem inversen Wert von 2 (= 1/2) gewichtet. Für unser Beispiel sind die Gewichte also:

1/RX für X
1/RY für Y
1/3 für die Dummies von Q
1/2 für die Dummies von P
1/2 für die Dummies von A
1/2 für die Dummies von B

Die Unähnlichkeit zwischen zwei Klassifikationsobjekten i und j wird nun berechnet nach:

$$
\begin{aligned}
CITY(i,j) = {} & Abs(X_i - X_j)/RX + ABS(Y_i - Y_j)/RY + \\
& \Sigma Abs(Q_{1i} - Q_{1j})/3 + \\
& \Sigma Abs(P_{1i} - P_{1j})/2 + \\
& \Sigma Abs(A_{1i} - A_{1j})/2 + \\
& \Sigma Abs(B_{1i} - B_{1j})/2 +
\end{aligned}
$$

Das entsprechende SPSS-X Programm hat folgende Struktur:

```
COMPUTE R1=Zahlenwert
COMPUTE R2=Zahlenwert
COMPUTE X=X/R1
COMPUTE Y=Y/R2
DO REPEAT DQ = DQ1 TO DQ4/ WERT = 1 TO 3
COMPUTE DQ = 0
IF (Q GE WERT) DQ = 1./2.
END REPEAT
DO REPEAT DP = DP1 TO DP4/ WERT = 1 TO 4
COMPUTE DP = 0
IF (P = WERT) DP = 1./3.
END REPEAT
DO REPEAT DA = DA1 TO DA3/ WERT = 1 TO 3
COMPUTE DA = 0
IF (A = WERT) DA = 1
END REPEAT
DO REPEAT DB = DB1 TO DB2/ WERT = 1 TO 2
COMPUTE DB = 0
IF (B = WERT) DB = 1
END REPEAT
CLUSTER X Y DP1 TO DP4 DQ1 TO DQ3 DA1 TO DA3 DB1 TO DB2
  /MEASURE = BLOCK
  /....
```

Bisher liegen kaum Befunde über die Brauchbarkeit des Unähnlichkeitsmaßes von Gower vor. Fehlende Werte könnten sehr elegant durch die Methode des paarweisen Ausscheidens gelöst werden. Steht dieses Verfahren nicht zur Verfügung, wird man in den meisten Fällen auf die Methode des fallweisen Ausscheidens zurückgreifen müssen, da die Anwendungsvoraussetzungen für die Methode der Mittelwertsubstitution nicht erfüllt sind.

Übungsaufgabe 17:

a) Berechnen Sie im nachfolgenden Beispiel den Informationsradius!

	A1	A2	A3
Kl.**Objekt** 1	.50	.30	.20
Kl.**Objekt** 2	.00	.10	.90

b) Berechen Sie im nachfolgenden Beispiel das Distanzmaß nach Gower:

	Kl.objekte	
Kl.merkmal	1	2
A	A1	A3
B	B1	B1
C	C1	C4
XI	3	4
X2	5	6

A und B sind nominale Klassifikationsmerkmale mit 3 Ausprägungen, C ist ein ordinales Klassifikationsmerkmal mit 4 Ausprägungen. XI und X2 sind quantitative Klassifikationsmerkmale mit den Spannweiten 5.0 für XI und 7.0 für X2.

4.3 Zusammenfassung

In diesem Abschnitt wurden die Transformation von Klassifikationsmerkmalen und die Behandlung fehlender Werte dargestellt. Dabei zeigten sich sehr deutlich die Grenzen von SPSS·X. Allerdings ist auch in anderen Statistiksoftwarepaketen, wie BMDP, SAS und CLUSTAN die Behandlung fehlender Werte unbefriedigend gelöst. In keinem der drei angeführten Programmsystemen ist die Methode des paarweisen Ausscheiden möglich. Dieses Defizit hat u.a. den Autor veranlaßt, selbst ein Programm zur Clusteranalyse zu schreiben (Bacher 1988). Dieses Programm wurde in das Statistikprogrammsystem ALMO (Holm 1988) integriert und steht dort zur Verfügung (15). Ferner wurde ersichtlich, daß vergleichende Untersuchungen über die Brauchbarkeit der Methoden der Transformation und der Behandlung fehlender Werte weitgehend fehlen. In dieser Richtung sind intensive Anstrengungen erforderlich, um dem Praktiker Entscheidungshilfen anbieten zu können.

5. Anmerkungen

Das Zustandekommen dieser Arbeit verdanke ich vor allem der unendlichen Geduld meiner Frau Marion, die das Manuskript kritisch durchgelesen und dadurch eine zu starke formale Abhandlung verhindert hat. Der Dank gebührt darüber hinaus den Initiatoren und Organisatoren des Quantkurses, der jedes Jahr im Spätsommer in Salzburg stattfindet, sowie den Studenten, die an diesem teilnahmen. Schließlich sei auf Seiten des Herausgebers Herrn Dr. W. H. Schröder und Herrn Dr. R. Metz gedankt sowie Waltraud Kannonier, die die Endfassung des Manuskripts redigierte.

(1) Die Geburtenrate wird i.d.R. nach der Formel

$$\frac{\text{Anzahl der Geburten in einer Periode}}{\text{mittlere Bevölkerung in der Periode}}$$

berechnet. Die mittlere Bevölkerung erhält man indem die Bevölkerungsgröße am Beginn und am Ende der Periode gemittelt wird. Die mittlere Bevölkerung ist folglich:

$$\text{mittlere Bev. in einer Periode} = \frac{\text{Bev. am Beginn} + \text{Bev. am Ende}}{2}$$

Zur Berechnung spezifischer Geburtenraten vgl. z.B. Costas (1985: 21-47).

(2) Die auf die Haushalte aggregierten Daten sind im Anhang dokumentiert. Frau Dr. Ursula Walter sei an dieser Stelle für das Überlassen der Daten herzlich gedankt.

(3) Diese ursprünglichen Ausprägungen waren.
1 Haushaltsvorstand (HV)
2 Hausfrau (HF)
3 Kind des HVehepaares
4 Kind des HV alleine
5 Kind der HF alleine
6 Kind von Sohn/Tochter
7 Kind von Inwohner
8 Ziehkind, Kind von Verwandten
9 Eltern des HVehepaares
10 Geschwister des HVehepaares

11 Ehegatten eines Kindes
12 Gesinde
13 Inwohner
14 Gesinde und Inwohner
15 Hilfskraft
16 abwesendes Kind des HV alleine
17 abwesendes Kind der HF alleine
18 abwesendes Kind des HVehepaares
19 abwesende Geschwister, Verwandte
20 Abwesendes Kind von Kindern
21 Kind von Geschwistern
22 Kind von Inwohnern
23 Kind von Gesinde
24 zusätzlicher Verwandter
25 Ehegatte von Geschwistern
26 Ehegatte von zusätzlichen Verwandten

(4) Dieses und die nachfolgenden Beispiele wurden während der Quantkurse im Spätsommer 1986 und 1987 im Rechenzentrum der Universität Salzburg auf einer VAX unter dem Betriebssystem VMS V4.3 gerechnet und vom Autor auf der PC-Version SPSS/PC+ von SPSS-X nachgerechnet. Auf die Eingabe mit SPSS-PC+ wird im folgenden aber nicht weiter eingegangen.

(5) Eine Ultrametrik liegt dann vor, wenn für drei Klassifikationsobjekte A, B und C gilt:

$$\ddot{u}(A,B) \leq \max(\ddot{u}(A,C),\ddot{u}(C,B)),$$

wobei ü(..)ein Unähnlichkeitsmaß ist. Die Gleichung besagt z.B., daß die Entfernung zwischen zwei Punkten A und B kleiner oder gleich dem Maximum der Entfernungen von A zu C und von C zu B ist. Sie ist z.B. in einem gleichschenkligen Dreieck erfüllt.

(6) Unähnlichkeitsmaße erfüllen die Dreiecksungleichung, wenn gilt:

$$\ddot{u}(A,B) \leq \ddot{u}(A,C) + \ddot{u}(C,B),$$

wobei ü(..) ein Unähnlichkeitsmaß ist. Diese Gleichung besagt z.B. (im Unterschied zur Ultrametrik), daß die direkte Entfernung zwischen A und B kleiner oder gleich der indirekten Entfernung ist, wenn B von A über C erreicht wird.

(7) vgl. Übungsaufgabe 5c.

(8) Für p = r liefert die Minkowskimetrik Distanzmaße, die die Dreiecksungleichung erfüllen. Darüber hinaus sind alle aus der Minkowskimetrik abgeleiteten Unähnlichkeitsmaße translationsinvariant.

Das bedeutet, daß zu jedem Klassifikationsmerkmal ein konstanter Term addiert werden kann, ohne daß sich dadurch die berechneten Unähnlichkeitsmaße ändern. Dagegen gilt die Eigenschaft der Skaleninvarianz nicht: Werden die Klassifikationsmerkmale mit einem Skalar (eine Zahl ungleich 0) multipliziert, verändern sich die berechneten Unähnlichkeitsmaße.

Die Metrikeigenschaft (Dreiecksungleichung) der Korrelation CORR(A,B) zwischen zwei Klassifikationsobjekten A und B und des Cosinus COS(A,B) hat in der Literatur zu heftigen Diskussionen geführt (vgl. Anderberg 1973: 113-114). Der Korrelationskoeffizient und der Cosinus besitzen eingeschränkte Metrikeigenschaften. In den transformierten Merkmalsausprägungen ist die Dreiecksungleichung erfüllt, aber nicht in den ursprünglichen Merkmalsausprägungen.

Für Ähnlichkeitsmaße lautet die Dreiecksungleichung (Späth 1980: 16) allgemein:

$$[ä(A,B) + ä(B,C)]ä(A,C) \geq ä(A,B)ä(B,C),$$

wobei ä(..) ein Ähnlichkeitsmaß ist.

(9) Mittelwerte sind dann repräsentativ für eine Verteilung, wenn diese eingipfelig und annähernd symmetrisch ist. Besitzt beispielsweise eine Verteilung zwei Gipfel, die mit gleichen Häufigkeiten auftreten, so ist das Auftreten des Mittelwerts sehr unwahrscheinlich.

(10) Gewichtung bei der Berechnung der Minkowskimetrik bedeutet, daß die absoluten Abweichungen von zwei Klassifikationsobjekten i und j in jedem Klassifikationsmerkmal l gewichtet wird mit:

$$g_l[Abs(X_{il} - X_{jl})]^p$$

Die Gewichtung der Klassifikationsmerkmale stellt dagegen eine Operation dar, die unmittelbar auf die Klassifikationsmerkmale selbst angewendet wird. Sie ist definiert als:

$$h_l X_{il}$$

$$[Abs(h_l X_{il} - h_l X_{jl})]^p = h_l{}^p[Abs(X_{il} - X_{jl})]^r.$$

Es gilt nun für diese gewichteten Klassifikationsmerkmale:

$$[Abs(h_l X_{il} - h_l X_{jl})]^p = h_l{}^p[Abs(X_{il} - X_{jl})]^r.$$

Setzen wir

ist die Identität beider Operationen erwiesen.

(11) Diese Eigenschaft folgt unmittelbar aus der Translationsinvarianz der Minkowskimetrik (vgl. Anmerkung 8).

(12) Der entscheidende Unterschied besteht darin, daß bei der Faktorenanalyse explizit Meßfehler in dem Modellansatz aufgenommen werden. Dadurch sind die Faktorwerte (Werte der Klassifikationsobjekte in den Faktoren) nicht mehr identifiziert. Bei der Hauptkomponentenanalyse dagegen stellen die »Faktorenwerte« (Wert der Klassifikationsobjekte in den Hauptkomponenten) Linearkombinationen der beobachteten Klassifikationsmerkmale dar und sind deshalb identifiziert.

(13) Bei der Anweisung INITIAL werden alle Eigenwerte und Eigenvektoren ausgegeben, während bei der Anweisung EXTRACTION nur die Eigenwerte und Eigenvektoren ausgegeben werden, die in die weitere Analyse einbezogen werden.

(14) Die Rotationsverfahren werden z.B. ausführlich in Arminger (1979) beschrieben.

(15) Das Programm**System** ALMO ist als Großrechnerversion und als PC-Version für ATARI-Computer verfügbar. Eine DOS-Version wird ab Spätsommer 1989 zur Verfügung stehen.

6. Literaturverzeichnis

Anderberg, M.R., 1973: Cluster Analysis for Applications. New York *Sc* London

Andrews, H.C., 1972: Introduction to mathematical techniques in pattern recognition. New York, London, Sydney *Sc* Toronto

Arminger, G., (1978): Faktorenanalyse. Stuttgart

Bacher, J., 1987: Missing data and measurement error in hierarchical cluster analysis: some simulation results. Nürnberg

Bacher, J., 1988: Clusteranalyse. S. 453-469 in: Holm, K.: ALMO Statistik-System. Linz

Bardeleben, H., 1987: CONCLUS. Gießen

Blaschke,G. *Sc* G. Liesegang, 1977: Die Klassifizierung von Nachfragekurven zur Verbesserung der kurzfristigen Absatzprognose in einem Betrieb mit modeabhängigem Produktionsprogramm. S. 37 - 53 in: Späth, H., (Hg.): Fallstudien Cluster Analyse. München & Wien

Blashfield, R.G. u.a., 1982: Validating a cluster analytic solution. S. 167 - 177 in: Hudson, H.C., (Hg.): Classifying social data. San Francisco, Washington *Sc* London

Bock, H., 1974: Automatische Klassifikation. Göttingen

Costas, l.,1985: Grundlagen der Wirtschafts - und Sozialstatistik. Frankfurt am Main

Fahrmeier, L., W. Häußler *Sc* G. Tutz, 1984: Diskriminanzanalyse. S. 301 - 370 in: Fahrmeier, L. *Sc* A. Hamerle, (Hg.): Multivariate statistische Verfahren. Berlin *Sc* New York

Fox, J.. 1982: Selective aspects of measuring resemblance for taxonomy. S. 172 - 151 in: Hudson, H.C., (Hg.): Classifying social data. San Francisco, Washington *Sc* London

Gower, J.C., 1971: A general coefficient of similarity and some of its propertics. Biometrics (23), S. 857 - 871

Hartigan, J.A., 1975: Clustering algorithms. New York, London, Sydney & Toronto

Holm, K. 1988: ALMO Statistik-System. Linz

Hudson, H.C., (Hg.): Classifying social data. San Francisco, Washington & London

Jardine, N. *Sc* R. Sibson, 1971: Mathematical taxonomy. London, New York, Sydney *Sc* Toronto

Kaufman, R.L., 1985: Issues in Multivariate Cluster Analysis: Some Simulation Results. Sociological Methods & Research (13:4), S. 467-487

Kaufmann, H. & H. Pape, 1984: Clusteranalyse. S. 371 - 473 in: Fahrmeir, L. & A. Hamerle, (Hg.): Multivariate statistische Verfahren. Berlin & New York

Klecka, W.R., 1984: Discriminant analysis. Beverly Hills & London

Mezzich, J.E., 1982: Comparing cluster analytic methods. S. 152-166 in: Hudson, H.C., (Hg.): Classifying social data. San Francisco, Washington & London

Schlosser, O., 1976: Einführung in die sozialwissenschaftliche Zusammenhangsanalyse. Reinbek bei Hamburg

Sixtl, F., 1982: Meßmethoden der Psychologie. 2. Auflage. Weinheim Basel

Sneath, P.H.A. & R.R. Sokal, 1973: Numerical taxonomy. San Francisco

Sodeur, W., 1974: Empirische Verfahren zur Klassifikation. Stuttgart

Späth, H., 1977: Fallstudien Cluster-Analyse. München *Sc* Wien

Späth, H., 1980: Cluster analysis algorithms for data reduction and classification of objects. New York

SPSS Inc., 1985: SPSS statistical algorithms. Chicago

SPSS Inc., 1986: SPSS-X User's guide. 2nd edition. Chicago

Vogel, F., 1975: Probleme und Verfahren der numerischen Klassifikation. Göttingen

Historical Social Research, Vol. 14 — 1989 — No. 2, 6-167

7. Anhang

7.1 Famiiiale Haushaltsdaten

1.Variable = HNR
2.Variable = AKERNF
3.Variable = AVERW
4.Variable = AINW
5.Variable = AGESIN

1	1	0	0	3
2	3	3	1	0
3	5	0	0	0
4	3	2	0	0
5	3	0	0	0
6	8	1	1	0
7	4	0	0	0
8	1	0	0	0
9	3	0	1	0
10	7	0	0	0
11	4	2	0	1
12	6	2	0	0
13	5	2	0	0
14	4	3	2	2
15	3	0	0	0
16	1	3	1	2
17	5	0	0	0
18	2	0	0	0
19	3	4	0	0
20	2	4	0	0
21	4	**0**	1	**0**
22	4	**0**	**0**	0
23	4	2	1	0
24	8	3	1	2
25	2	0	1	0
26	7	1	0	0
27	6	3	0	0
28	3	6	0	0
29	4	3	0	0
30	1	8	0	4

31	5	0	0	1
32	1	2	0	1
33	7	0	0	0
34	2	2	0	0
35	2	1	0	0
36	5	0	0	0
37	7	6	0	0
38	1	5	0	0
39	1	3	0	0
40	2	3	0	0
41	1	3	0	1
42	1	1	0	0
43	5	1	0	0
44	4	1	0	0
45	2	0	0	0
46	5	1	0	0
47	4	3	1	0
48	5	0	0	0
49	8	0	0	0
50	2	5	0	0
51	3	1	0	0
52	2	9	0	0
53	1	3	0	1
54	6	0	0	0
55	4	0	0	0
56	4	0	0	0
57	6	0	0	1
58	10	0	0	0
59	2	5	0	1
60	8	0	0	0
61	4	1	0	0
62	7	7	0	0
63	4	0	0	0
64	2	8	**0**	**0**
65	1	4	**0**	1
66	4	1	0	1
67	4	1	0	0
68	2	0	0	0
69	2	0	1	0
70	2	5	0	1
71	4	1	0	0
72	3	0	0	0
73	5	5	0	0

74	3	0	0	0
75	8	0	0	0
76	7	1	0	1
77	8	1	0	0
78	8	0	0	0
79	5	3	0	1
80	10	0	0	0
81	4	1	0	0
82	2	5	0	0
83	3	0	0	0
84	2	0	1	0
85	6	0	0	0
86	2	0	0	2
87	4	8	0	0
88	7	0	0	0
89	2	5	0	0
90	6	0	0	0
91	1	3	0	0
92	5	2	0	0
93	6	0	0	4
94	3	2	0	0
95	6	1	0	3
96	3	0	0	0
97	5	8	0	0
98	6	5	0	0
99	5	2	0	2
100	10	1	0	0
101	7	1	0	0
102	1	2	0	0
103	5	0	0	1
104	1	5	0	0
105	6	7	0	0
106	5	2	0	0
107	7	0	0	3
108	5	1	0	3
109	4	4	0	0
110	4	3	0	0
111	4	3	0	0
112	1	8	1	1
113	3	0	0	0
114	3	0	1	0
115	3	2	0	0
116	9	0	0	1

117	3	0	1	0
118	2	6	1	0
119	5	3	0	0
120	3	6	0	0
121	5	4	0	0
122	4	0	0	5
123	8	2	0	0
124	4	0	0	0
125	5	2	0	1
126	2	2	0	0
127	5	0	0	0
128	7	0	0	0
129	6	3	0	2
130	1	0	0	0
131	8	1	0	3
132	4	0	0	0
133	2	0	0	2
134	5	0	0	0
135	5	0	0	0
136	5	0	0	0
137	3	0	0	0
138	12	2	0	0
139	2	5	0	0
140	4	9	0	0
141	8	3	0	0
142	6	1	0	2
143	6	2	1	2
144	4	2	0	0
145	3	7	0	0
146	1	3	0	0
147	1	4	0	0
148	2	1	0	0
149	1	0	0	0
150	4	1	0	1
151	4	0	0	Ü
152	2	0	0	0
153	2	0	0	0
154	2	0	0	0
155	1	1	0	0
156	1	1	0	0
157	2	0	0	2
158	5	0	0	2
159	2	0	0	0

7.2 Die SPSS-X Prozedur QUICK CLUSTER

Die Grundidee der SPSS-X Prozedur QUICK CLUSTER (vgl. SPSS Inc. 1985) soll anhand eines einfachen Beispiels beschrieben werden. Gegeben ist folgende Klassifikationsdatenmatrix:

		Kl.merkm. X1	X2
Kl.objekte	1	3	0
	2	2	0
	3	1	0
	4	0	1
	5	0	2
	6	0	3

Der Benutzer muß die **Anzahl der Cluster** vorgeben, wenn diese nicht der Voreinstellung von 2 Clustern entspricht. Ferner kann zwischen folgenden Optionen gewählt werden:

1. Auswahl der Startwerte der Clustermittel werte (Initial Cluster Centers) nach der Reihenfolge (NOINITIAL) oder systematisch nach der Trennschärfe der Klassifikationsobjekte (INITIAL).
2. Aktualisierung der Clustermittelwerte (Voreinstellung) oder keine Aktualisierung der Clustermittelwerte (NOUPDATE).

Die erste Option bestimmt die Auswahl der Startwerte der Clustermittelwerte:

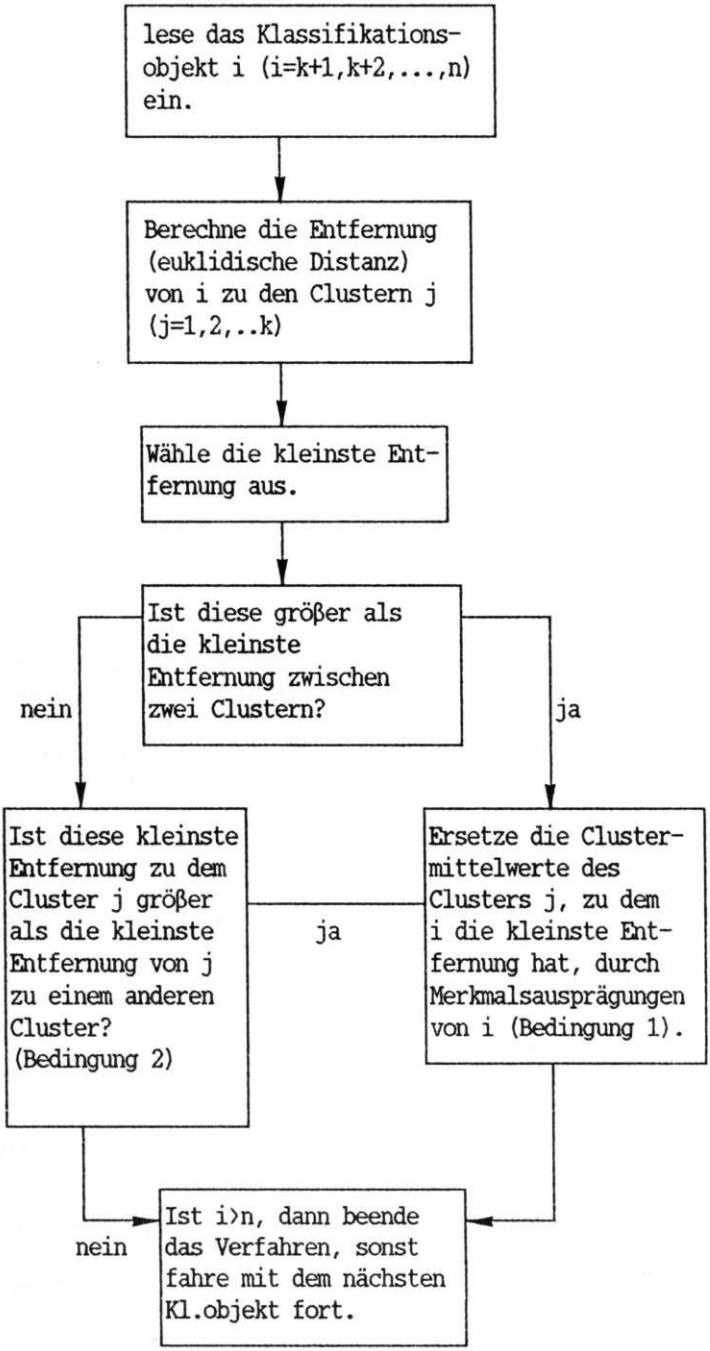
Reihenfolgeauswahl: Das erste Klassifikationsobjekt bildet das erste Cluster, das zweite das zweite Cluster, usw..

Systematische Auswahl: Bei k Clustern bilden die ersten k Klassifikationsobjekte die k Cluster. Diese Cluster und ihre Mittelwerte werden nun folgendermaßen geändert:

(Ablaufschema siehe folgende Seite):

Für die fiktiven Daten führt dieser Algorithmus zu folgenden Startclustermittel werten:

1. Clustermittelwert = Merkmalsausprägungen des 1. Klassifikationsobjektes
2. Clustermittelwert = Merkmalsausprägungen des 2. Klassifikationsobjektes


lese das Klassifikations-objekt i (i=k+1,k+2,...,n) ein.
Berechne die Entfernung (euklidische Distanz) von i zu den Clustern j (j=1,2,..k)
Wähle die kleinste Entfernung aus.
Ist diese größer als die kleinste Entfernung zwischen zwei Clustern?
nein
ja
Ist diese kleinste Entfernung zu dem Cluster j größer als die kleinste Entfernung von j zu einem anderen Cluster? (Bedingung 2)
ja
Ersetze die Clustermittelwerte des Clusters j, zu dem i die kleinste Entfernung hat, durch Merkmalsausprägungen von i (Bedingung 1).
nein
Ist i>n, dann beende das Verfahren, sonst fahre mit dem nächsten Kl.objekt fort.

Die minimale Entfernung zwischen zwei Clustern ist in diesem Fall identisch mit der Entfernung von Cluster 1 und Cluster 2 (= 1.0).

Einlesen des Klassifikationsobjektes 3:

Das Klassifikationsobjekt 3 besitzt mit 1.0 die kleinste Entfernung zum Cluster 2. Die erste Bedingung ist nicht erfüllt und da nur zwei Cluster vorliegen die zweite Bedingung ebenfalls nicht. Das 3. Klassifikationsobjekt führt also zu keiner Änderung der Startclustermittelwerte.

Einlesen des Klassifikationsobjektes 4:

Das Klassifikationsobjekt 4 besitzt mit 1.44 die kleinste Entfernung zum Cluster 2. Die erste Bedingung ist erfüllt. Die Startclustermittelwerte werden durch die Merkmalsausprägungen des 3. Klassifikationsobjektes ersetzt. Sie betragen nun 0 und 1. Die kleinste Entfernung zwischen zwei Clustern erhöht sich dadurch auf 3.16.

Einlesen des Klassifikationsobjektes 5:

Das Klassifikationsobjekt 5 besitzt mit 1.0 die kleinste Entfernung zum Cluster 2. Beide Bedingungen sind nicht erfüllt, die Startclustermittelwerte ändern sich folglich nicht.

Einlesen des Klassifikationsobjektes 6:

Das Klassifikationsobjekt 6 besitzt mit 2.0 die kleinste Entfernung zum Cluster 2. Die erste Bedingung ist erfüllt und das Cluster 2 erhält die neuen Startclustermittelwerte 0 und 3.

Zusammenfassend ergeben sich also folgende Startclustermittelwerte:

		Auswahl nach Reihenfolge		systematische Auswahl	
		X1	X2	X1	X2
Cluster	1	3.	0.	3.	0.
	2	2.	0.	0.	3.

Soll keine Aktualisierung der Startclustermittelwerte durchgeführt werden, sind die Startclustermittelwerte mit den Klassifikationsmittelwerten (Classification Cluster Centers) identisch und die Klassifikationsobjekte werden jenem Cluster zugeordnet, zu dem die Entfernung minimal ist. Dieses Vorgehen ergibt folgendes Ergebnis:

Auf der linken Seite stehen die Ergebnisse bei einer Auswahl der Startwerte nach der Reihenfolge; auf der rechten Seite die bei einer systematischen Auswahl.

Classification Cluster Centers (Klassifikations-clustermittelwerte):

	X1	X2	X1	X2
Cluster 1	3.	0.	3.	0.
Cluster 2	2.	0.	0.	3.

Clusterzugehörigkeit:

Kl.objekt		
1	1	1
2	2	1
3	2	1
4	2	2
5	2	2
6	2	2

Final Cluster Centers (Zielclustermittelwerte):

	X1	X2	X1	X2
Cluster 1	3.0	0.0	2.0	0.0
Cluster 2	0.6	1.2	0.0	2.0

Aktualisierung der Clustermittelwerte

Bei der **Aktualisierung** werden die Clustermittel werte nach jeder Zuordnung eines Klassifikationsobjektes geändert. Das genaue Vorgehen ist folgendes: Nach Berechnung der Startclustermittelwerte besteht das Cluster c aus n_c Objekten und besitzt die Clustermittelwerte

$Xj_{,}$, x_{2c}.

Die Klassifikationsobjekte werden nun von vorne beginnend dem Cluster, **zu** dem sie die geringste Entfernung besitzen zugeordnet. Nach jeder Zu**Ordnung** werden die Clustermittelwerte neu berechnet. Wird das Klassifikationsobjekt j dem Cluster c **zugeordnet,** so betragen die neuen Clustermittel werte

$$X_{ic}' = (n_c X_{ic} + X_{ij})/(n_c + 1).$$

Die Clustergröße wird für das nächste Klassifikationsobjekt um 1 auf $n_c' = n_c + 1$ erhöht. Für die Berechnung der Entfernung der verbleibenden Klassifikationsobjekte zu dem Cluster c werden diese neuen Clustermittelwerte verwendet. Wird nun als nächstes Klassifikationsobjekt, das Klassifikationsobjekt k, dem Cluster c zugeordnet, betragen die neuen Clustermittel werte

$$X_{ic}'' = (n_c' X_{ic}' + X_{ik})/(n_c' + 1).$$

Die neue Clustergröße beträgt nach dieser Zuordnung $n_c'' = n_c' + 1$.

In unserem Beispiel ergeben sich z.B. für die Auswahl der Startclustermittelwerte nach der Reihenfolge folgende Klassifikationsclustermittelwerte:

Klassifikationsobjekt 1: wird dem 1. Cluster zugeordnet, neue Clustermittelwerte sind 3.0 und 0.0.

Klassifikationsobjekt 2: wird dem 2. Cluster mit einer Clustergröße von 5.0 zugeordnet. Die neuen Clustermittelwerte betragen nun (5*2.0 + 2.0)/6 = 2 für XI und (5*0 + 0)/6 *m* 0 für X2, die Clustergröße erhöht sich auf 6.0.

Klassifikationsobjekt 3: wird dem 2. Cluster zugeordnet. Die neuen Clustermittelwerte betragen nun (6*2.0 + 1.0)/7 = 1.857 für XI und (6*0 + 0)/7 = 0 für XI. Die Clustergröße erhöht sich auf sieben.

Klassifikationsobjekt 4: wird dem 2. Cluster zugeordnet. Die neuen Clustermittelwcrtc betragen nun (7*1.875 + 0)/8 « 1.625 für XI und (7*0 + 1)/8 = .125. Die Clustergröße erhöht sich auf acht.

Klassifikationsobjekt 5: wird dem 2. Cluster zugeordnet. Die neuen Clustermittelwerte betragen nun (8*1.625 + 0)/9 = 1.444 für XI und (8*0.125 + 2)/9 = 0.33 für X2. Die Clustergröße erhöht sich auf neun.

Klassifikationsobjekt 6: wird dem 2. Cluster zugeordnet. Die neuen Clustermittelwerte betragen nun (9*1.444 + 0)/9 = 1.299 für XI und (9*0.33 + 3)/10 - 0.59 für X2. Die Clustergröße erhöht sich auf 10. Die neuen Klassifikationsclustermittelwerte sind 1.299 für XI und 0.59

für X2. Auf ihrer Grundlage werden die Klassifikationsobjekte den Clustern zugeordnet, zu dem sie die geringsten Entfernungen besitzen.

Als Ergebnisse erhält man die Zielclustermittelwerte:

Zielclustermittelwerte:

	X1	X2		X1	X2
Cluster 1 (n=2)	2.50	0.00	(n=3)	2.0	0.0
Cluster 2 (n=4)	0.25	1.50	(n=3)	0.0	2.0

Auf der linken Seite stehen die Ergebnisse, wenn die Startwerte nach der Reihenfolge der Klassifikationsobjekte ausgewählt werden, auf der rechten Seite die Ergebnisse bei einer systematischen Auswahl der Startwerte der Clustermittelwerte.

In dem Beispiel führt ganz offensichtlich eine Auswahl der Startwerte der Clustermittel werte nach der Reihenfolge zu keiner befriedigenden Clusterlösung, unabhängig davon ob die Clustermittelwerte aktualisiert werden oder nicht. (Zur Kritik an dem QUICK CLUSTER-Verfahren vgl. Hartigan 1975: 74 - 83.) Die Ergebnisse hängen stark von der Reihenfolge der Cluster ab. Eine Änderung der Reihenfolge führt zu anderen Ergebnissen. Bei der Suche nach einer empirischen Klassifikation wird man deshalb die systematische Auswahl von Clustermittelwerten wählen. Die Aktualisierung der Clustermittelwerte hat nur einen starken Einfluß auf das Klassifikationsergebnis, wenn die Clustergrößen klein sind, da die Größe der Startcluster und die Startclustermittelwerte in die Aktualisierung der Cluster eingehen.

Mit QUICK CLUSTER kann neben einer empirischen Klassifikation auch eine sehr rudimentäre konfirmatorische Clustersnalyse vorgenommen werden. In diesem Fall können durch die Anweisung

INITIAL = (theoret. Clustermittelwerte des I.Clusters
theoret. Clustermittelwerte des 2.Clusters
...
theoret. Clustermittelwerte des n.Clusters)

die theoretischen Clustermittelwerte eingelesen werden. Zusätzlich muß auf eine Aktualisierung der Clustermittelwerte verzichtet werden. Das Verfahren ist deshalb rudimentär, da nicht die Möglichkeit besteht, nur bestimmte Clustermittelwerte theoretisch zu fixieren und die verbleiben-

den Clustermittelwerte empirisch schätzen zu lassen. Ein Programm das diese flexiblere Handhabung ermöglicht, wurde von Bardeleben (1987) entwickelt.

Für unsere familialen Haushaltsdaten führt die Prozedur QUICK CLUSTER für 3 Cluster zu folgenden Ergebnissen, wenn die Startwerte systematisch ausgewählt und aktualisiert werden:

Zeilclustermittelwerte:

	AKERNF	AVERW	ABIW	AGESIN
Cluster 1 (n=98)	3.18	1.01	0.13	0.41
Cluster 2 (n=28)	3.18	6.04	0.07	0.25
Cluster 3 (n=33)	7.48	0.88	0.09	0.61

Die Ergebnisse stimmen weitgehend mit denen des Complete-Linkage (vgl. Kapitel 2) überein. Das symmetrische Lambda beträgt 0.81.

In QUICK CLUSTER kann zusätzlich die Entfernung jedes Klassifikationsobjektes zu dem Cluster, dem es angehört, und die Entfernung der Clustermittelwerte zueinander ausgegeben werden. Die entsprechenden Anweisungen sind:

/PRINT - INITIAL CLUSTER ID(NKl.) DISTANCE ANOVA

Die Anweisung INITIAL bewirkt die Ausgabe der Startcluster - , der Klassifikationscluster - und der Zielclustermittelwerte. Durch die Anweisung CLUSTER wird die Clusterzugehörigkeit jedes Klassifikationsobjektes und dessen Entfernung zum Cluster, dem es angehört, ausgegeben. Durch die Spezifikation 1D(NKL.) wird den Klassifikationsobjekten der Name NKL zur Identifikation zugewiesen. NKL. muß dabei eine Stringvariable sein. Die Entfernung der Clustermittelwerte zueinander wird bei Verwendung des Befehls DISTANCE ausgegeben. Durch die Anweisung ANOVA wird eine einfache Varianzanalyse für die erzielte Clusterlösung durchgeführt. Dadurch kann der Beitrag der einzelnen Klassifikationsmerkmale zur Trennung der Cluster gemessen werden. Für die familialen Haushaltsdaten sind die entsprechenden Werte:

Ergebnisse der Varianzanalyse:

Kl.merkmale	MSQS zw. Cluster	DF	MSQS in. Cluster	DF	F	Prob
AKERNF	242.02	2	2.3016	156	105.2	.00
AVERW	295.38	2	1.5735	156	187.7	.00
AINW	.05	2	.1100	156	0.4	.64
AGESIN	.98	2	.8500	156	1.2	.32

MSQS zw. Clustern (Cluster MS in SPSS X) ist die mittlere Streuungsquadratsumme zwischen den Clustern, die sich durch Division der Streuungsquadratsumme mit den Freiheitsgraden ergibt. Sie gibt die Heterogenität zwischen den Clustern an. Die Anzahl der Freiheitsgrade von MSQS zw. Clustern ist immer gleich der Anzahl der Cluster minus 1. Die mittlere Streuungsquadratsumme innerhalb der Cluster (MSQS in. Cluster, Error MS in SPSSX) mißt dagegen die Heterogenität zwischen den Clustern. Den F Wert erhält man durch die Division von MSQS zw. Clustern mit MSQS in. Cluster. Unter der H_0-Hypothese, daß alle Clustermittelwerte identisch sind, und der Annahme, daß alle Klassifikationsobjekte Realisierungen einer Normalverteilung darstellen, besitzt F eine FVerteilung mit Dfl und Df2 Freiheitsgraden. Der Wert PROB gibt nun an, mit welchem Fehler diese H_0-Hypothese verworfen werden kann. In dem Beispiel der familialen Haushaltsdaten unterscheiden sich die Cluster ausschließlich durch die Klassifikationsmerkmale AKERNF und AVERW.

Historical Social Research, Vol. 14 — 1989 — No. 2, 6-167

8. Lösungen der Übungsaufgaben

1-a) quantitativ, b) nominal, c) ordinal und d) quantitativ

2) Struktur der Ausgangsdaten (fiktive Ausprägungen):

Person	BERUF	EINK	WFL
1	1	4	20,5
2	2	3	6,5
.	.	.	.
.	.	.	.
n	3	1	5,5

Struktur der Klassifikationsdatenmatrix:

Kl.objekte	Kl.merkmale	
Berufe	Dummies von EINK	durchschn. WFL
	DEINK1 DEINK2 DEINK3 DEINK4	
BERUF = 1		
2		
3		
.		
.		
.		
14		

SPSS-X Programm:

```
TITLE »Aggregation der Datei V.DAT über die Berufe«
COMMENT
COMMENT Definition und Einlesen der Datei
COMMENT
FILE HANDLE VDAT / NAME - »V.DAT«
GET FILE - VDAT
COMMENT
COMMENT Definition der Dummy-Variablen für Einkommen
COMMENT
```

```
COMPUTE DEINK1 -= 0
COMPUTE DEINK2 - 0
COMPUTE DEINK3 « 0
COMPUTE DEINK4 - 0
COMMENT
COMMENT Wertzuweisung zu den Dummies:
COMMENT DEINK(i) « 1, wenn EINK - i
COMMENT
IF (EINK EQ 1) DEINK1 « 1
IF (EINK EQ 2) DEINK2 - 1
IF (EINK EQ 3) DEINK3 - 1
IF (EINK EQ 4) DEINK4 - 1
COMMENT
COMMENT Definition des Ergebnisfiles AVZ.DAT für
COMMENT Aggregationsergebnisse
COMMENT
FILE HANDLE - AVZDAT / NAME - »AVZ.DAT«
COMMENT
COMMENT Beginn der Aggregation
COMMENT
AGGREGATE OUTFILE - AVZDAT
      /BREAK « BERUF
      /AEINK1 ,AEINK2,AEINK3,AEINK4,AWFL -
      MEAN(DEINK1,DEINK2,DEINK3,DEINK4,WFL)
```

3-a) nicht zulässig, da Familienstand nominales Meßniveau besitzt.

b) Aussage ist falsch, da sich z.B. die Unähnlichkeit zwischen zwei Klassifikationsobjekten in nominalen Klassifikationsmerkmalen mit unterschiedlichen Ausprägungen mit Hilfe der City Block-metrik problemlos berechnen läßt (vgl. Abschnitt 2.4.4). Vergleich-barkeit ist in diesem Fall immer gegeben, da die City-Blockmetrik immer Werte von 1 oder 0 annimmt.

4-a) Verschmelzungsschema für den Complete- und Single- Linkagc:

Complete-Linkage			Single-Linkage		
Schritt	Cluster	Niveau	Schritt	Cluster	Niveau
1	A,B	1.0	1	A,B	1.0
2	C,D	1.0	2	C,D	1.0
3	CD,E	4.0	3	AB,CD	2.0
4	AB,CDE	12.0	4	ABCD,E	4.0

Auf der Grundlage des Verschmelzungsschemas können die Dendrogramme leicht konstruiert werden.

b) nein, da Familienstand und Geschlecht nominale Klassifikationsmerkmale darstellen.

c) z.B. bei einer Klassifikation von Arten oder evolutionären Entwicklungen (s. Jardine *Sc* Sibson 1971: 127 - 166).

5-a) $CITY(1,2) = 9 + 10 + 15 + 14 = 48$

$EUCLID(1,2) = (9^2 + 10^2 + 15^2 + 14^2)^{1/2} = 24.5$

$COSINUS(1,2) = (10.1 + 24.2 + 18.3 + 16.2)/$
$((10^2 + 12^2 + 18^2 + 16^2)(1^2 + 2^2 + 3^2 + 2^2))^{1/2} =$
$120 / 824.18)^{1/2} = .985$

$POWER(1,2/3,1) = (9^3 + 10^3 + 15^3 + 14^3) - 7848$

b) z.B. wenn Güter aufgrund ihrer relativen Preisschwankungen klassifiziert werden sollen.

c) nein. Es gilt zwar für standardisierte Klassifikationsobjekte:

$$SEUCLID(i,j) = \Sigma(Z_{ik} - Z_{jk})^2 = \Sigma Z_{ik}^2 + \Sigma Z_{jk}^2 - 2\Sigma Z_{ik}Z_{jk}$$
$$= 1 + 1 - 2COS(i,j) = 2(1 - COS(i,j))$$

Bei nichtstandardisierten Klassifikationsobjekten dagegen wird i.d.R. ΣX_{ik}^2 und ΣX_{jk}^2 ungleich 1 sein und die Identität gilt deshalb nicht.

6-a) Merkmalsausprägungen für Anteilswerte:

	Vor Normierung mit Anteilsw.						Nach Normierung auf Anteilswerte					
Kl.obj.	Kl.merkmale				MW	S	Kl.merkmale				MW	S
1	1	1	0	0	.5	2	.50	.50	.00	.00	.25	1
2	2	2	0	0	1.0	4	.50	.50	.00	.00	.25	1
3	2	2	1	1	1.5	6	.33	.33	.33	.33	.25	1
4	3	3	1	1	2.0	8	.38	.38	.38	.38	.25	1*
5	4	4	1	1	2.5	10	.40	.40	.10	.10	.25	1

*) genauer Wert = .375

b) Merkmalsausprägungen nach Mittelwertzentrierung:

	Vor Normierung mit Mittelw.					Nach Normierung mit Mittelwerten					
Kl.obj.	Kl.merkmale				MW	Kl.merkmale				MW	S
1	1	1	0	0	.5	.5	.5	-0.5	-0.5	.0	0
2	2	2	0	0	1.0	1.0	1.0	-1.0	-1.0	.0	0
3	2	2	1	1	1.5	.5	.5	-0.5	-0.5	.0	0
4	3	3	1	1	2.0	1.0	1.0	-1.0	-1.0	.0	0
5	4	4	1	1	2.5	1.5	1.5	-1.5	-1.5	.0	0

Die Mittelwertzentrierung bewirkt also - wie die Berechnung der Anteilswerte - daß die absolute Höhe der Klassifikationsobjekte in den Klassifikationsmerkmalen eliminiert wird. Allerdings führen beide Operationen zu unterschiedlichen Unähnlichkeitsmaßen. Betrachten wir z.B. die City-Blockmetrik zwischen dem Klassifikationsobjekt 1 und 2, so beträgt diese 0.0 für die Anteilswerte und 2.0 bei einer Mittelwertzentrierung.

Zwischen beiden Operationen besteht folgende Beziehung: Die Anteilswerte sollen mit $Z_{ik} = X_{ik}/(mX_{i\cdot})$ bezeichnet werden (m = Anzahl der Klassifikationsmerkmale, $X_{i\cdot}$ = Mittelwert des Klassifikationsobjektes i in den Klassifikationsmerkmalen), die mittelwertzentrierten Beobachtungen mit $U_{ik} = X_{ik} - X_{i\cdot}$. Es gilt nun:

$$Z_{ik} = a + b_i X_{ik}$$
$$\text{mit } a = -1/m$$
$$b_i = 1/mX_{i\cdot}$$

c) SPSS-X Programm für Mittelwertzentrierung:

```
TITLE »Mittelwertzemrierung der Kl.Objekte«
COMMENT
COMMENT Definition und Einlesen der Daten
COMMENT
FILE HANDLE AFAMD / NAME -= »AFAM.DAT«
GET FILE « AFAMD
COMMENT
COMMENT Berechnen der Mittelwerte der Kl.objekte
```

```
COMMENT mit der Funktion MEAN
COMMENT
COMPUTE MKORJ -MEAN(AKERNF,AVERW,AINW,AGESIN)
COMMENT
COMMENT Durchführen der Mittelwertzentrierung
COMMENT
COMPUTE ZKERNF - AKERNF · MKOBJ
COMPUTE ZVERW = AVERW · MKOBJ
COMPUTE ZINW - AINW · MKOBJ
COMPUTE ZGES1N - AGES1N · MKOBJ
COMMENT
COMMENT Ende der Mittelwertzentrierung, es können
COMMENT weitere SPSS-X Befehle folgen
COMMENT
```

7) SPSS-X Programm:

```
TITLE »Lösung der Übungsaufgabe 7«
COMMENT
COMMENT Definiton und Einlesen der Daten
COMMENT
FILE HANDLE PDAT / NAME - »P.DAT«
GET FILE - PDAT
COMMENT
COMMENT Rekodierung des Kl.merkmales REL, damit
COMMENT einheitlicher Zahlenbereich
COMMENT
RECODE REL (9=4)
COMMENT
COMMENT Das Kl.merkmal REL wird in nominale Dummies
COMMENT aufgelöst und mit 1/2 gewichtet, damit ein
COMMENT heitlicher Wertebereich der City-Block
COMMENT metrik mit ordinalen Kl.merkmalen
COMMENT (s. Distanzmaß von Gower).
COMMENT
DO REPEAT DREL - DREL1 TO DREL4/WERT = 1 TO 4
COMPUTE DREL= 0
IF (REL EQ WERT) DREL - 1/2
END REPEAT
COMMENT
COMMENT Das Kl.merkmal GRUND wird in ordinale Dummies
COMMENT aufgelöst und mit 1/3 gewichtet (3 = Anzahl
COMMENT der Ausprägungen minus 1 ).
COMMENT
DO REPEAT DGRUND=DGRUND1 TO DGRUND2/WERT= 1 TO 4
COMPUTE DGRUND = 0
IF (GRUND GE WERT) DGRUND= 1/3
END REPEAT
COMMENT
COMMENT Das Kl.merkmal HERK wird in ordinale Dummies
COMMENT aufgelöst und mit 1/2 gewichtet (2 •= Anzahl der
COMMENT Ausprägungen minus 1).
```

```
COMMENT
DO REPEAT DHERK - DHERK1 TO DHERK3/WERT = 1 TO 3
COMPUTE DHERK-0
IF (HERK GE WERT) DHERK- 1/2
END REPEAT
COMMENT
COMMENT Ende der Übungsaufgabe, es können weitere
COMMENT SPSS·X Befehle folgen.
COMMENT
```

8) Bedeutung der neuen Klassifikationsmerkmale:

AALT - durchschnittliches Alter je Beruf
AREL1 - Anteil röm.kath. Personen je Beruf
AREL2 - Anteil evang. Personen je Beruf
AREL3 - Anteil jüd. Personen je Beruf
DHERK1 — kumulierter gewichteter Anteil von Einheimischen je Beruf
DHERK2 — kumulierter gewichteter Anteil von Einheimischen und Nahwanderer je Beruf
DHERK3 — kumulierter gewichteter Anteil von Einheimischen, Nah· und Fernwanderer je Beruf
DGRUND1 - kumulierter gewichteter Anteil von Grundbebesitz von 0 bis unter 1 ha je Beruf
DGRUND2 — kumulierter gewichteter Anteil von Grundbesitz von 0 bis unter 2 ha je Beruf
DGRUND3 — kumulierter gewichteter Anteil von Grundbesitz von 0 bis unter 5 ha je Beruf
DGRUND4 — kumulierter gewichteter Anteil von Grundbesitz von 0 bis über 5 ha je Beruf

9) Die reproduzierten Unähnlichkeitsmatrizen sind:

	A	B	C	D	E
A	0				
B	1	0			
C	12	12	0		
D	12	12	1	0	
E	12	12	4	4	0

Complete-Linkage

	A	B	C	D	E
A	0				
B	1	0			
C	2	2	0		
D	2	2	1	0	
E	4	4	4	4	0

Single-Linkage

Die Matrixkorrelation beträgt folglich:

i	C_i	S_i	C_i^2	S_i^2	$C_i S_i$
1	1	1	1	1	1
2	12	2	144	4	24
3	12	2	144	4	24
4	12	4	144	16	48
5	12	2	144	4	24
6	12	2	144	4	24
7	12	4	144	16	48
8	1	1	1	1	1
9	4	4	16	16	16
10	4	4	16	16	16
S	82	26	898	82	226

$S_{CC} = 898 - 82*82/10 = 225.6$
$S_{SS} = 82 - 26*26/10 = 14.4$
$S_{CS} = 226 - 26*82/10 = 12.8$
$CORR(C,S) = 12.8/(14.4*225.6)^{1/2} = 0.22$

10) X = Baverage-Linkage
Y = Single-Linkage
Fehler ohne X = 2
Fehler mit X = 2
Fehler ohne S = 35
Fehler mit S = 34
h_{XY} = 0 (Baverage --> Single)
h_{YX} = .029 (Single --> Baverage)
sym.h_{XY} = 0.028

II) 1. Programm:

```
TITLE -ÜBUNGSAUFGABE 11«
COMMENT
COMMENT Definition und Einlesen der Ausgangsdaten
COMMENT
FILE HANDLE F DAT / NAME - F. DAT
GET FILE = FD AT
COMMENT
COMMENT Rekodierung des Kl.merkmales REL, damit
COMMENT durchgehender Wertebereich
COMMENT
RECODE REL<9=4)
```

```
COMMENT
COMMENT Auflösung von REL in Dummies
COMMENT
DO REPEAT DREL - DREL1 TO DREL4 / WERT - 1 TO 4
COMPUTE DREL - 0
IF (REL EQ WERT) DREL = 1
END REPEAT
COMMENT
COMMENT Auflösung von FAMST in Dummies
COMMENT
DO REPEAT DFAM - DFAM1 TO DFAM4 / WERT - 1 TO 4
COMPUTE DFAM - 0
IF (FAMST EQ WERT) DFAM = 1
END REPEAT
COMMENT
COMMENT Berechnen der 3 Clusterlösungen.
COMMENT Die Ergebnisse werden in den Variablen C3 und
COMMENT S3 zwischengespeichert
COMMENT
CLUSTER DREL1, DREL2, DREL3, DREL4, AFAM1, AFAM2, AFAM3,
        AFAM4
      /MEASURE - BLOCK
      /METHOD = COMPLETE (C) SINGLE (S)
      /PLOT = NONE
      /PRINT = NONE
      /SAVE - CLUSTERS)
COMMENT
COMMENT Abspeichern der Ergebnisse des Complete Linkage auf
COMMENT die Datei C.DAT
COMMENT
FILE HANDLE = CD AT / NAME = »CD AT«
AGGREGATE OUTF1LE = CD AT
      /BREAK = C3
      /AREL1,AREL2,AREL3,AREL4,AFAM1, AFAM2, AFAM3,
      AFAM4 -
      MEAN(DREL1 ,DREL2,DREL3,DREL4, DFAM 1, DFAM2,
      DFAM3, DFAM4)
COMMENT
COMMENT Abspeichern der Ergebnisse des Single auf die
COMMENT Datei S.DAT, die Ausprägungen von S3 werden
COMMENT zuvor rekodiert.
COMMENT
RECODE S3 (1 = 4) (2=5) (3=6)
FILE HANDLE = S DAT / NAME = »S.DAT«
AGGREGATE OUTFILE = S DAT
      /BREAK = S3
      /AREL1 ,AREL2,AREL3,AREL4,AFAM 1, AFAM2, AFAM3,
      AFAM4 =
      MEAN(DREL1,DREL2,DREL3,DREL4,DFAM1, DFAM2,
      DFAM3, DFAM4)
```

2. Programm:

```
TITLE »Übungsaufgabe 11, 2.Teil«
COMMENT
COMMENT Definition der Eingabefiles
COMMENT
FILE HANDLE CDAT / NAME «  »CDAT«
FILE HANDLE SDAT / NAME - »S.DAT«
COMMENT
COMMENT Aneinanderfügen der Files
COMMENT
ADD FILES FILE - CDAT / RENAME «= (C3 - S3)
      /FILE - SDAT
COMMENT
COMMENT Berechnen der City-Blockmetrik zwischen
COMMENT den Clusterlösungen in der Prozedur CLUSTER
COMMENT
CLUSTER AREL1,AREL2,AREL3,AREL4, AFAM1, AFAM2, AFAM3,
        AFAM4
      /MEASURE - BLOCK
      /PRINT * DISTANCE
      /PLOT - NONE
```

12)
```
TITLE »Übungsaufgabe 12« COMMENT
COMMENT Definition und Einlesen der Datei FDAT
COMMENT
FILE HANDLE FDAT / NAME - »F.DAT«
GET FILE = FDAT
COMMENT
COMMENT Rekodierung von REL, damit einheitlicher
COMMENT Wen ebereich
COMMENT
RECODE REL (9=4)
COMMENT
COMMENT Berechnen einer gleichvert. Zufallsvariablen
COMMENT
COMPUTE FEHLER - UNIFORM(IOO)
COMMENT
COMMENT Berechnen der Dummy-Variablen FEHLK (0=kein
COMMENT Fehler, 1= Fehler)
COMMENT
IF (FEHLER LE 10.0) FEHLK = 1
COMMENT
COMMENT Berechnen einerneuen gleichvert. Zufallszahl
COMMENT zur Bestimmung der Kategorie, zu dem falsch
COMMENT zugeordnet wird
COMMENT
COMPUTE FEHLER = UNIFORM(IOO)
COMMENT
COMMENT Bestimmung des Fehlers ( 0 bis 33% - 1, 33 bis
COMMENT 67% = 2 und 67 bis 100% = 3)
COMMENT
```

```
IF (FEHLER LE 33) AA=1
IF (FEHLER GT 33 AND FEHLER LE 67) AA-2
IF (FEHLER GT 67) AA « 3
COMMENT
COMMENT Rekodierung von AA, wenn REL= 1, dann soll AA-= 1
COMMENT zu AA = 4 werden, usw.
COMMENT
IF (FEHLK EQ 1 AND REL EQ 1 AND AA EQ 1) AA-4
IF (FEHLK EQ 1 AND REL EQ 2 AND AA EQ 2) AA-4
IF (FEHLK EQ 1 AND REL EQ 3 AND AA EQ 3) AA-4
COMMENT
COMMENT AA wird gleich REL gesetzt, falls kein Fehler
COMMENT auftritt
COMMENT
IF (FEHLK EQ 0) AA-REL
COMMENT
COMMENT Zuweisung der fehlerhaften bzw. gültigen Ausprägung zu
COMMENT REL
COMMENT
COMPUTE REL-AA
COMMENT
COMMENT Für das Klassifikationsmerkmal FAMST wird
COMMENT analog verfahren
COMMENT
```

13-a) siehe Text

b)
```
TITLE »Übungsaufgabe 13b«
COMMENT Definition und Einlesen der Daten
COMMENT
FILE HANDLE FDAT / NAME - »F.DAT«
GET FILE FDAT
COMMENT
COMMENT Auflösung der Kl.merkmale in Dummies
COMMENT
RECODE REL (9=4)
DO REPEAT DREL = DREL1 TO DREL4 / WERT = 1 TO 4
COMPUTE DREL = 0
IF (REL EQ WERT) DREL = 1
END REPEAT
DO REPEAT DFAM = DFAM1 TO DFAM4 / WERT = 1 TO 4
COMPUTE DFAM = 0
IF (FAMST EQ WERT) DFAM = 1
END REPEAT
COMMENT
COMMENT Durchführen der Clusteranalyse und
COMMENT Zwischenspeichern der Ergebnisse
COMMENT
CLUSTER DREL1 ,DREL2,DREL3.DREL4, AFAM1, AFAM2, AFAM3,
        AFAM4
      /MEASURE = BLOCK
      /METHOD - COMPLETE (C)
```

```
        /PRINT - NONE
        /PLOT - NONE
        /SAVE - CLUSTER (4)
COMMENT
COMMENT Durchführen der Diskriminanzanalyse, in diese
COMMENT dürfen nicht alle Dummy-Variablen aufgenommen
COMMENT werden, da sonst lineare Abhängigkeiten ent-
COMMENT stehen. Die Anzahl der Dummies eines Kl.merk-
COMMENT males, die in die Diskriminanzanalyse einbe-
COMMENT zogen werden können, ist Anz. d. Dummies - 1
COMMENT
DISCRIMINANT GROUPS - C4(l,4)
        /VARIABLES - DREL1, DREL2, DREL3, DFAM1, DFAM2,
        DFAM3
        /METHOD « DIRECT
        /PRIORS - [Zahlenwerte]
        /STATISTICS - ALL
```

c) Bei der Berechnung der Anzahl der Diskriminanzfunktionen sind die in die Analyse einbezogenen Dummies (und nicht die Anzahl der ursprünglichen Klassifikationsmerkmale) entscheidend. Folglich gibt es in 13b drei Diskriminanzfunktionen (Anzahl der Gruppen - 1), da die Anzahl der Dummies größer 3 ist.

14-a) Rechenschema zur Berechnung der City-Blockmetrik bei fehlenden Werten:

Paare	Absolute Differnz in AKERNF	ÄVERW	Anw	AGESIN	s	Resk. faktor	CITY
HH1,HH2	2	3	*	3	8	4/3	10.7
HH1,HH3	4	*	0	3	7	4/3	9.3
HH1,HH4	2	2	0	*	4	4/3	5.3
HH1,HH5	2	0	0	3	5	4/4	5.0
HH2,HH3	2	*	*	0	2	4/2	4.0
HH2,HH4	0	1	*	*	1	4/2	2.0
HH2,HH5	0	3	*	0	3	4/3	4.0
HH3,HH4	2	*	0	*	2	4/2	2.0
HH3,HH5	2	*	0	0	2	4/3	2.7
HH4,HH5	0	2	0	*	3	4/3	4.0

b) nein, da kein gemeinsames Meßmodell

c) Der Mittelwert MX1 ist in diesem Fall nicht definiert und erhält SPSS-X intern den fehlenden Wert SYSMIS. Dieser Wert wird in den folgenden COMPUTE-Anweisungen auch den Klassifikationsmerkmalen XI, X2 und X3 zugewiesen.

15-a) Durch die Anweisung COMPUTE S1GMA1 = wert wird der Variablen SIGMA1 die Standardabweichung von XI zugewiesen, durch die Anweisung COMPUTE SIGMA2 - wert der Variablen SIGMA2 die Standardabweichungen von X2, usw.
In den anschließenden COMPUTEAnweisungen werden die Klassifikationsmerkmale varianznormalisiert.

b) Wir wollen zunächst zeigen, daß die Varianznormalisierung und ^Transformation bei der City-Blockmetrik zu identischen Ergebnissen führt. Seien $Z_{il} = X_{il}/\sigma_l$ die varianznormalisierten Merkmalsausprägungen und $U_{il} = (X_{il} \cdot \mu_l)/\sigma_l$ die Ztransformierten Merkmalsausprägungen, dann ist:

$$\begin{aligned} Abs(U_{il} \cdot U_{jl}) &= Abs(X_{il}/\sigma_l \cdot \mu_l/\sigma_l \cdot X_{jl}/\sigma_l + \mu_l/\sigma_l) \\ &= Abs(Z_{il} \cdot Z_{jl}) \end{aligned}$$

Aus dieser Identität folgt unmittelbar die Behauptung.
Für den Cosinus dagegen führen beide Operationen zu unterschiedlichen Ergebnissen: Gegeben seien z.B. die beiden Klassifikationsmerkmale XI und X2 mit den Mittelwerten μ_1 - 0 und μ_2 = 1 und den Standardabweichungen o, = 1.0 und σ_2 = 2.0 und die beiden Klassifikationsmerkmale 1 und 2 mit den Ausprägungen $X_{11} = 10, X_{12} = 1, X_{21} = 1$ und $X_{22} = 2$. Für die varianznormalisierten Klassifikationsmerkmale beträgt der Cosinus 0.53 und für die Ztransformierten -1.0. Die Behauptung gilt somit nicht allgemein.

16-a) quantitativ, da ansonsten keine gewichtete Summenbildung (= Berechnung von Linearkombinationen) möglich ist.

b) 1.Eigenwert - $.45^2 + .88^2 + .77^2 + .13^2 + .23^2 + .12^2$ = 1.64 (27%),
2.Eigenwert - $.33^2 + .20^2 + (-.23)^2 + .69^2 + .60^2 + .72^2$ = 1.56 (26%)

c) Als Gewichte können unmittelbar die Eigenvektoren verwendet werden. Das entsprechende SPSS-X Programm ist:

```
TITLE »Übungsaufgabe 16«
COMMENT
COMMENT Definition und Einlesen der Daten
COMMENT
FILE HANDLE AFAMD / NAME - »AFAM.DAT«
```

```
GET FILE - AFAMD
COMMENT
COMMENT Berechnen der 1. Hauptkomponente (Werte aus Tabelle 4.2-5)
COMMENT
COMPUTE Hl -= -.77*AKERNf + .73*AVERW + .31*AINW -
                .13*AGESIN
COMMENT
COMMENT Berechnen der 2. Hauptkomponente
COMMENT
COMPUTE H2 « .02*AKERNF - .13*AVERW + .69*AINW -
                .76*AGESIN
COMMENT
COMMENT Berechnen der 3. Hauptkomponente
COMMENT
COMPUTE H3 - -.02*AKERNF -I- .35*AVERW - .63*AINW +
                .62*AGESIN
COMMENT
COMMENT Berechnen der 4. Hauptkomponente
COMMENT
COMPUTE H4 « .64*AKERNF + .57*AVERW + .19*AINW -
                .09*AGESIN
```

17-a) $1(1/1 + 2) = .50*\log_2(0.50/.50) + .30*\log_2(.30/.40) + .20*\log_2(.20/1.10) \approx .383$

$1(2/1 + 2) = .00*\log_2(.00/.50) + .10*\log_2(.10/.40) + .90*\log_2(.90/1.10) = .540$

$1(1,2) = (1(1/1+2) + 1(2/1+2))/2 = .46$

b) Rechenschema zur Berechnung des Distanzmaßes von Gower:

	A1	A2	A3	B1	B2	B3	C1	C2	C3	C4	XI	X2
1	1	0	0	1	0	0	1	0	0	0	3	5
2	0	0	1	1	0	0	1	1	1	1	5	6
abs. Differenz.	2			0			3				2	1
Gewichte	1/2			1/2			1/3				1/5	1/7
gew. abs. Differenz	1			0			1				.4	.14

GOWER(1,2) = 1 + 0 + 1 + .4 + .14 = 2.54